TRACE: TRANSPARENT REASONING THROUGH ALIGNED CONCEPT EXPLANATIONS IN CLINICAL REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

When an offline reinforcement learning (RL) policy recommends a treatment that differs from observed clinical practice, the disagreement alone cannot reveal whether the departure reflects clinically useful personalization or limitations in the data or learning objective. We introduce Transparent Reasoning through Aligned Concept Explanations (TRACE), a framework that explains and directly compares a learned RL policy with an estimated behavior policy fitted to recorded treatment actions. TRACE uses a shared encoder and sparse autoencoder to define a common concept space, then conditions a shared LLM-based explainer on the patient state and policy-relevant concepts while withholding the selected action. Teacher supervision and proximal policy optimization promote hidden-action recovery, concept consistency, state grounding, and resistance to action leakage. We evaluate TRACE on three MIMIC-III/IV tasks, including renal replacement therapy, noninvasive ventilation, and vasopressin initiation, where the policies disagree on 11.9%–33.4% of held-out states. TRACE achieves the highest hidden-action recovery accuracy and macro-F1 among the evaluated explanation methods for both policies. On disagreement states, each policy's action is recovered more accurately from its own explanation than from the other policy's explanation. Concept perturbations reduce learned-policy recovery across all tasks, while policy-side interventions show that cited concepts are more influential than uncited or random concepts, most clearly for RRT. These results show that TRACE enables faithful, concept-grounded comparison of learned and behavior policies, providing a basis for pre-deployment auditing and targeted clinical review of disagreements.

1 INTRODUCTION

Optimizing sequences of clinical decisions to improve patient outcomes has become an important area of application for offline reinforcement learning (RL), including sepsis management, ventilation support, and kidney transplantation (Komorowski et al., 2018; Peine et al., 2021; Hua et al., 2022; Tang et al., 2026; Yousuf et al., 2026). By learning treatment policies from previously collected patient trajectories, offline RL avoids prospective trial-and-error interaction with patients, which may be impractical or unsafe (Levine et al., 2020). A central challenge in evaluating such policies arises when a learned policy recommends an action that differs from observed clinical practice. Such departures may reflect clinically meaningful personalization (Murphy, 2003; Yazzourh et al., 2026), but they may also arise from limitations in the logged data or learning objective, including limited support, unobserved confounding, and reward misspecification (Fujimoto et al., 2019; Namkoong et al., 2020; Skalse et al., 2022; Luo et al., 2024b). The differing actions alone do not explain why the learned policy departs from practice. This motivates explainable RL methods that identify which aspects of the patient's state contributed to the learned recommendation and thereby support closer examination of the disagreement (Milani et al., 2024; Tonekaboni et al., 2019).

Interpreting such a disagreement requires explanations of both the learned policy's recommendation and the observational practice from which it diverges, which we model using an estimated behavior policy. Fitted to treatment actions recorded in the observational data, this policy reflects population-level treatment tendencies rather than the decision process of any individual clinician. For each policy, a useful explanation should identify the patient-state information contributing to its action

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.

(Tonekaboni et al., 2019) and faithfully reflect the policy's computation (Bauer & Michalowski, 2025). To enable direct comparison when the policies disagree, the explanations should additionally express decision-relevant information in a shared vocabulary (Miller, 2019).

Existing methods do not fully address this comparative setting. Most explainable RL methods explain policies in isolation (Milani et al., 2024; Bekkemoen, 2024; Luo et al., 2024a; Melloni & Zingoni, 2026), whereas prior policy-comparison methods primarily characterize differences in observable actions, trajectories, or preferences rather than the internal bases of those differences (Amitai et al., 2024). Natural language offers an intuitive interface for communicating decision-relevant information, but a large language model (LLM) can generate a plausible clinical rationale without reflecting the information actually used by the policy (Turpin et al., 2023; Lyu et al., 2024; Matton et al., 2025). Moreover, independently generated explanations need not share a vocabulary, and successful recovery of a hidden policy action is insufficient by itself: the explanation may reveal the action through correlational or stylistic cues rather than through the internal features that drove the policy (Hase et al., 2020; Jiang et al., 2024; Parcalabescu & Frank, 2024; Yang et al., 2026). We therefore ask: *Can natural-language explanations faithfully capture the decision basis of each policy and distinguish those bases when the policies disagree on the same patient state, using aligned, mechanistically testable concepts?*

To address this gap, we introduce **Transparent Reasoning through Aligned Concept Explanations (TRACE)**, a framework for faithful, policy-specific explanation and direct comparison of learned and estimated behavior policies (Figure 1). TRACE maps each patient state to a common representation using a shared encoder, with policy-specific branches producing the learned-policy and behavior-policy decisions. After training, we hold the encoder and both branches fixed and fit a sparse autoencoder (SAE) (Makhzani & Frey, 2014; Huben et al., 2024; Gao et al., 2025) to decompose the encoder representations into a shared dictionary of interpretable concepts. This dictionary gives concepts common semantics across policies, while policy-specific concept selection and decision effects characterize how each policy uses them. For each patient state and policy, a shared LLM-based explainer receives the state, a target-policy indicator, and a small set of policy-relevant concept cards, while the selected action itself is withheld. The explainer is initialized using teacher-generated explanations and subsequently optimized with proximal policy optimization (PPO) (Schulman et al., 2017) to preserve decision-relevant information while maintaining concept consistency, state grounding, and resistance to direct action leakage. We first assess whether an independent judge can recover the hidden policy action from the explanation. Beyond action recovery, we examine whether the explanation retains information specific to its source policy and whether its cited concepts participate in the frozen policy's computation.

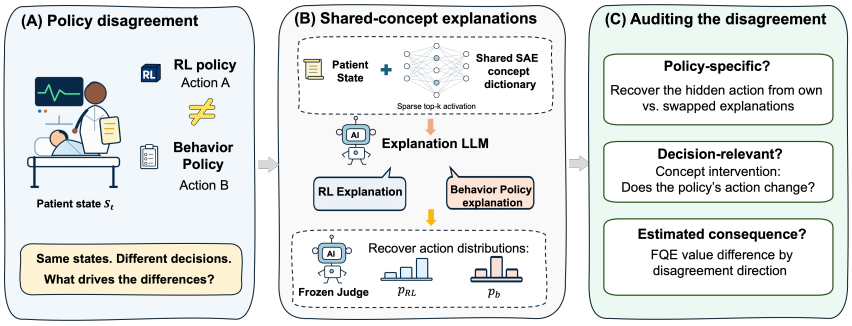


Figure 1: **Overview of TRACE. (A)** The learned RL and estimated behavior policies may choose different actions for the same patient state. **(B)** TRACE uses a shared SAE concept dictionary to generate action-blind, policy-specific explanations, evaluated through hidden-action recovery by a frozen judge. **(C)** Cross-policy swapping and concept interventions audit policy specificity and decision relevance, while FQE provides model-based context across disagreement directions.

We evaluate TRACE on three offline clinical RL tasks constructed from MIMIC-III and MIMIC-IV (Johnson et al., 2016a; 2024): renal replacement therapy (RRT) modality selection, noninvasive ventilation (NIV) switching, and vasopressin initiation in septic shock, where the learned and estimated behavior policies disagree on 11.9%–33.4% of held-out states. Across all three tasks, TRACE achieves the highest hidden-action recovery accuracy and macro-F1 among the evaluated explana-

tion methods for both policies; cross-policy comparisons and targeted interventions further indicate that its explanations preserve policy-specific information and depend on decision-relevant concepts. Because neither explanation preference nor retrospective FQE establishes which action is clinically superior, we position TRACE as a framework for pre-deployment policy auditing rather than automated clinical adjudication.

Our contributions are threefold: **(1) An aligned concept space for cross-policy comparison.** TRACE decomposes shared encoder representations into a sparse concept dictionary, fixing common concept semantics across policies while policy-specific concept selection and decision effects enable direct comparison on the same patient state. **(2) Action-blind, policy-specific natural-language explanations.** A shared LLM-based explainer generates explanations from patient states and policy-specific concept evidence without observing the selected action. It is optimized for decision relevance, concept consistency, state grounding, and resistance to action leakage. **(3) A disagreement-centered, two-level evaluation of faithfulness.** Across three critical-care RL tasks, cross-policy swapping tests source-policy specificity, while interventions on concept cards and frozen policy representations test whether selected and cited concepts affect explanation recovery and policy decisions. Paired explanation comparisons and direction-stratified FQE further contextualize disagreement patterns without serving as clinical validation.

1.1 Related Work

Clinical RL oversight and policy comparison. Clinical offline RL is learned from observational trajectories, making evaluation sensitive to limited support, unobserved confounding, and uncertainty in off-policy estimation (Gottesman et al., 2019; Namkoong et al., 2020; Luo et al., 2024b; Otten et al., 2024). Consequently, disagreement with recorded care establishes neither clinical superiority nor unreliability. Prior work compares policies through disagreement-based behavioral summaries (Amitai & Amir, 2022; Amitai et al., 2024), contrastive explanations of agent preferences (Gajcin et al., 2021), and counterfactual clinical audits (Ren & Liao, 2026). These methods illuminate where policies differ, but do not jointly explain the internal decision evidence used by a learned policy and an estimated behavior policy in a shared concept space. TRACE addresses this gap by comparing the two policies on matched patient states.

Concept-based interpretability and aligned representations. Concept-based methods describe model behavior through human-interpretable representations (Kim et al., 2018; Poeta et al., 2025). Concept bottleneck models rely on predefined concepts (Koh et al., 2020; Delfosse et al., 2024), whereas post-hoc and label-free approaches discover concepts without concept annotations (Oikarinen et al., 2023; Schut et al., 2025). Sparse autoencoders (SAEs) decompose internal representations into sparse latent features that can be inspected and intervened upon (Huben et al., 2024; Bricken et al., 2023; Gao et al., 2025; Marks et al., 2025; Basu et al., 2026). Representation-similarity methods compare networks at an aggregate level (Kornblith et al., 2019; Klabunde et al., 2025), while universal SAEs and sparse crosscoders seek shared feature coordinates across models or layers (Thasarathan et al., 2025; Minder et al., 2025). TRACE brings this shared-coordinate perspective to policy comparison: a common encoder and SAE dictionary fix concept semantics, while policy-specific concept selection and decision effects characterize how each frozen policy uses those concepts and ground paired natural-language explanations.

Faithful natural-language policy explanations. Prior work has generated natural-language rationalizations of agent behavior (Ehsan et al., 2018; Hayes & Shah, 2017; Admoni et al., 2026), while recent methods use LLMs to verbalize surrogate models, feature attributions, or other XRL signals (Milani et al., 2024; Zhang et al., 2025; Belouadah et al., 2025; Kim et al., 2026; Gyevnár et al., 2026); see Milani et al. (2024) for a broader review. Most closely related, Yang et al. (2026) train an Explanation LLM partly through rewards for recovering hidden policy decisions. Such recovery measures simulatability—whether an explanation contains information predictive of the policy action—but does not by itself establish that the explanation reflects the policy's internal computation and may be inflated by superficial or label-revealing cues (Hase & Bansal, 2020; Hase et al., 2020; Jiang et al., 2024). TRACE therefore supplements action recovery with cross-policy swapping to test source-policy specificity, concept-card perturbations to test the explainer's dependence on supplied evidence, and feature removal to test effects on frozen policy decisions.

2 METHOD

TRACE generates policy-specific natural-language explanations for a learned RL policy and an estimated behavior policy defined over the same state and action spaces. It represents both policies through a shared state encoder, decomposes the resulting representation into a common concept dictionary, and identifies how each policy uses those concepts for a given patient state. A shared Explanation LLM then translates the policy-specific concept evidence into natural language without observing the selected action or its scores. We next formalize each component of the framework.

2.1 POLICY DISAGREEMENT

We consider a learned RL policy π_{RL} and an estimated behavior policy π_b over the same state and action spaces. For a state s, let $a^{\mathrm{RL}}(s)$ denote the action returned by the learned policy's complete decision rule and $a^b(s) = \arg\max_a \pi_b(a \mid s)$ the most likely action under the behavior policy. We define the disagreement set as

$$\mathcal{D}_{\mathrm{dis}} = \{s : a^{\mathrm{RL}}(s) \neq a^b(s)\}. \tag{1}$$

The behavior policy is fitted to treatment actions in the observational data and therefore represents population-level treatment patterns conditional on the observed state. After policy learning, both policies are held fixed throughout concept construction and explanation training. TRACE characterizes the decision-relevant concept evidence used by each policy, allowing their decisions to be examined in a common representation.

2.2 SHARED POLICY REPRESENTATION

To represent the two policies in a shared space, we use a common state encoder $h(s) = E(s) \in \mathbb{R}^d$ with separate decision branches, so that

$$\pi_{\mathrm{RL}}(\cdot \mid s) = F_{\mathrm{RL}}(h(s)), \qquad \pi_b(\cdot \mid s) = F_b(h(s)), \tag{2}$$

where the shared representation $h(s)$ is learned jointly during policy training. The learned-policy branch is optimized using the offline RL objective, whereas the behavior-policy branch is trained using cross-entropy on observed actions. Both objectives update the shared encoder, placing the two policies in a common representation space while allowing them to learn different decision rules.

The learned-policy branch F_{RL} implements the complete decision rule, including any admissible-action constraints. The behavior-policy branch F_b consists of a residual adapter, $h_b(s) = h(s) + g_b(h(s))$, followed by a prediction head (Jayaraman et al., 2024). The adapter and behavior head are updated only by the behavior objective. After training, both policies and the shared encoder are frozen, and $h(s)$ is used to train the SAE.

2.3 CONCEPT CONSTRUCTION AND SELECTION

Shared concept dictionary. For each cohort we build one dictionary shared by both policies, ranking sparse features mainly by their contribution to reconstructing the representation and secondarily by their contributions to Q values, action imitation, and clinician prediction. The final dictionary is selected by cohort-specific fidelity criteria. To interpret a feature we compare clinical variables across its low, mid, and high activation groups on training subjects and verify that the directions hold on held-out validation subjects. The descriptions are cohort relative, with the mid group as a reference. Features without a reliable interpretation keep their slot as explicitly uncertain cards with no name or direction, which the teacher may not cite as evidence (Supplementary Section S4.2).

Concept selection. For each state s and target policy $p \in \{\mathrm{RL}, b\}$ we choose a small set of active concepts $C_p(s)$, favoring concepts that contribute more to the policy's representation and to why it selects one action over the alternatives: through Q values and the rule that restricts allowed actions for the RL policy, and through the predicted action preference for the behavior policy, with a penalty on concepts that repeat the same clinical information. The selected action guides this step but is never shown to the Explanation LLM. Each card carries a name, an activation level, a cohort-relative description, and the patient's relevant measurements.

2.4 POLICY COMPARISON

The shared dictionary provides a common basis for comparing the two policies. For a state s, let $z_{\mathrm{RL}}(s) = f_{\mathrm{SAE}}(h(s))$ and $z_{\mathrm{B}}(s) = f_{\mathrm{SAE}}(h_b(s))$, where a given coordinate refers to the same dictionary concept in both representations. Because the dictionary is fitted on h, z_{RL} is always is always evaluated within its training domain. We interpret z_{B} only when the dictionary reconstructs h_b with sufficient fidelity, defined as a fraction of variance unexplained no greater than 0.25. This criterion is satisfied for the vasopressin cohort, permitting direct comparison of concept activations, but not for RRT or NIV, where the adapter shifts the representation by more than half its norm. For these two cohorts, we instead compare each concept's *decision effect*, defined as the change in a policy branch's decision margin after removing that concept from h. Because the complete decision branch is rerun after removal, this comparison does not require the SAE to reconstruct h_b (Supplementary Section S3.3). We compare activation differences where reconstruction is reliable and decision-effect differences otherwise, using matched patient states and stratifying the analyses by policy agreement and disagreement.

2.5 TEACHER SUPERVISION AND SHARED PPO TRAINING

For each benchmark, we train a single Explanation LLM to explain both the learned RL policy and the behavior policy. Let $a^p(s)$ denote the decision of policy $p \in \{\mathrm{RL}, b\}$ at state s, and let $x_p(s) = (s, p, C_p(s))$ contain the patient state, a policy indicator, and the corresponding policy specific concepts. The policy indicator specifies which policy is being explained, allowing the explainer to generate explanations for both policies sharing all model parameters.

Algorithm 1 TRACE: shared concept construction and explainer training.

Require: Frozen policies $\pi_{\mathrm{RL}} = F_{\mathrm{RL}}(E(\cdot))$, $\pi_b = F_b(E(\cdot))$; training states $\mathcal{D}$; Teacher LLM; frozen reader q_R; base explainer π_{θ_0}
Ensure: Shared explainer π_θ for $p \in \{\mathrm{RL}, b\}$
Stage 1: Shared concept dictionary
Fit a Top-k SAE on $h(s) = E(s)$ by minimizing $\mathcal{L}_{\mathrm{SAE}}$
Select the dictionary by policy fidelity and assign concept names
Stage 2: Policy-specific concept evidence
for each $s \in \mathcal{D}$ and $p \in \{\mathrm{RL}, b\}$ **do**
  Compute policy action $a^p(s)$
  Select policy-relevant concepts $C_p(s)$ using activation and decision contribution
  Set $x_p(s) \leftarrow (s, p, C_p(s))$
end for
Stage 3: Supervised initialization
for each (s, p) **do**
  Generate $e^\star \leftarrow \textsc{Teacher}(s, C_p(s), a^p(s))$
  Retain $e^\star$ only if grounding and leakage checks pass
end for
$\pi_{\mathrm{SFT}} \leftarrow \mathrm{SFT}\big(\pi_{\theta_0}, \{(x_p(s), e^\star)\}\big)$
$\pi_\theta \leftarrow \pi_{\mathrm{SFT}}$
Stage 4: Shared PPO
for $t = 1, \ldots, N$ **do**
  Sample pair-aligned (s, p) and generate $e \sim \pi_\theta(\cdot \mid x_p(s))$
  Compute reader reward $R_p(s, e)$ using frozen q_R (Eq. 3)
  Update θ with PPO objective $\mathcal{J}(\theta)$ and KL penalty to π_{SFT}
end for
Select the checkpoint with highest development recovery accuracy

Teacher supervision. We first construct supervised training data using a Teacher LLM. Given the patient state, policy specific concepts, and the target policy decision, the teacher generates an explanation that connects the relevant concepts to the observed clinical evidence without revealing the action. We then use these teacher generated explanations to supervise fine tuning (SFT) of the explainer, providing an initial policy aware explanation model before PPO optimization.

Shared Policy Fine-Tuning. Building on the SFT initialization, we optimize the shared explainer with PPO on examples from both policies. A frozen reader q_R scores whether the explanation helps recover the target decision. For policy p, define

$$R_p(s, e) = \lambda_r \sum_j \Big[\log q_R\big(a^p(s) \mid s, p, e_{\le j}\big) - \log q_R\big(a^p(s) \mid s, p, e_{<j}\big) \Big] + \lambda_c r_{\mathrm{con}}\big(e, C_p(s)\big) + \lambda_g r_{\mathrm{gnd}}(e, s) - \lambda_\ell r_{\mathrm{leak}}(e). \tag{3}$$

The first term credits increases in the reader's probability of the target action as the explanation unfolds, at sentence boundaries. The other three are deterministic checks on the finished text:

$r_{\text{con}} \in \{-1, 0, 1\}$ sums an indicator for opening with the selected-concept framing in cohort-relative language and one for citing at least two patient measurements listed on the named cards, minus one; $r_{\text{gnd}} \in \{-1, 0, 1\}$ is $+1$ when every number matches the reference text within a display-rounding tolerance, -1 when any does not, and 0 when none is stated; $r_{\text{leak}} \in \{0, 1\}$ fires when the text names an action or intervention from a cohort lexicon copied from the teacher's acceptance rule, uses a reader label letter, or identifies which policy it explains. Implement details can be found in Supplementary Sections S2. We jointly train on both RL and behavior policy examples:

$$\mathcal{J}(\theta) = \sum_{p \in \{\text{RL}, b\}} \rho_p \mathbb{E}_{s,e|p} \left[R_p(s, e) - \beta \log \frac{\pi_\theta(e \mid x_p(s))}{\pi_{\text{SFT}}(e \mid x_p(s))} \right], \qquad \sum_p \rho_p = 1. \tag{4}$$

This shared objective produces a single explainer that preserves policy specific evidence for both policy targets.

3 EXPERIMENTS

In this section, we aim to demonstrate that TRACE can faithfully explain both learned and behavior policies, preserves their policy specific differences, and grounds its explanations in decision relevant concepts. We further use paired explanation comparisons to examine how differences between the two policies are reflected in their explanations.

Clinical benchmarks. We study three representative clinical cohorts covering distinct settings: renal replacement therapy (RRT) modality selection (Xu et al., 2026), noninvasive ventilation (NIV) switching (Feng et al., 2023; Lijović et al., 2025), and optimal vasopressin initiation in septic shock (OVISS) (Kalimouttou et al., 2025), using the MIMIC-III and MIMIC-IV databases (Johnson et al., 2016a;b; 2023; 2024; Goldberger et al., 2000). The three cohorts contain 2,095 RRT patients, 2,145 NIV ICU stays, and 3,102 vasopressin patients, respectively. We use RRT as a representative case for more detailed analysis, where the action space contains three actions: (1) No RRT, (2) intermittent hemodialysis (IHD), and (3) continuous renal replacement therapy (CRRT). The per-action results can be found in Supplementary Table S3, and Figure 3 shows the judge's preference between the explanations of the two policies on the held-out RRT test states. We conduct the same analyses for the NIV and OVISS cohorts, and the results can be found in Supplementary Tables S4 and Figure S1.

Baselines. We compare against baselines spanning prompting, supervised fine-tuning, reinforcement learning, and alternative policy-explanation frameworks. (1) **Direct LLM**: a frozen LLM explains from the state alone. (2) **CoT**: zero-shot Chain-of-Thought (CoT) (Kojima et al., 2022) from the state and concept cards, and CoT + SFT (Wei et al., 2022). (3) **MCTS Refinement** (Belouadah et al., 2025): inference-time tree search over candidate explanations. (4) **SFT**: the explainer fine-tuned on teacher explanations from the state alone, without concepts or PPO. (5) **PPO + BT**: our Concept-SFT initialization and inputs, with the reward replaced by a Bradley–Terry preference model (Hunter, 2004). (6) **BC-LLM** (Feng et al., 2025): Bayesian concept bottlenecks with LLM priors, verbalized. (7) **Policy Distillation + LLM** (Zhang et al., 2025): a decision-tree surrogate whose path is verbalized by an LLM. (8) **RLAIF + CNF** (Yang et al., 2026): distributional AI-feedback rewards modeled with continuous normalizing flows (CNF).

Policies. We use Batch-Constrained Q-learning (BCQ) (Fujimoto et al., 2019) as the fixed learned policy, with cohort-specific reward functions following the corresponding source studies (Xu et al., 2026; Feng et al., 2023; Kalimouttou et al., 2025). These rewards combine terminal survival with intermediate physiological signals reflecting organ dysfunction for RRT, respiratory recovery for NIV, and hemodynamic response for OVISS (Supplementary Section S1.2).

Models and Training. For each cohort, we construct a shared teacher-generated training dataset for all trainable explainers. For each patient state in the training split, we construct two training examples, one for the RL policy and one for the behavior policy. Each example contains the state, the target policy indicator, the selected concept cards, and a teacher-generated explanation. GPT 5.6-Terra (OpenAI, 2026a) serves as the Teacher LLM and generates explanations conditioned on the target action. Only explanations that pass deterministic grounding and leakage checks are retained.

All trainable methods use the same teacher-generated dataset, with concept cards removed for state-only variants. By default, we use Llama-3.1-8B-Instruct as the Explanation LLM, initialized with SFT and further refined using shared PPO, while Qwen3-8B remains frozen as the training-time reader that provides the PPO reward. The final judge plays the role of the clinician in our pipeline: it reads the patient state and an explanation, predicts the policy's action, and compares the two policies' explanations of the same patient. To reduce the bias of any single judge, Claude Opus 5 (Anthropic, 2026a) scores every evaluation, Claude Sonnet 5 (Anthropic, 2026b) re-scores action recovery for five arms per cohort (Supplementary Section S4.6), and GPT-6 Astra (OpenAI, 2026b), from a second model family, independently judges the paired explanation preference.

3.1 Explainer Performance Comparisons and Ablations

Table 1: Accuracy and macro-F1 across three cohorts (per method: RL row above behavior row).

Method	RRT		NIV		OVISS	
	Acc.	F1	Acc.	F1	Acc.	F1
Direct LLM	0.437	0.295	0.500	0.374	0.678	0.627
	0.492	0.403	0.622	0.367	0.675	0.806
Zero-shot CoT	0.438	0.291	0.512	0.339	0.720	0.659
	0.457	0.382	0.638	0.289	0.862	0.926
MCTS Refinement	0.373	0.258	0.528	0.365	0.500	0.493
	0.428	0.371	0.647	0.468	0.807	0.893
BC-LLM	0.487	0.341	0.502	0.392	0.617	0.615
	0.608	0.512	0.592	0.399	0.684†	0.682
Policy distill. + LLM	0.412	0.302	0.537	0.365	0.683	0.673
	0.423	0.355	0.633	0.371	0.888	0.941
SFT	0.697	0.639	0.545	0.420	0.765	0.730
	0.873	0.859	0.578	0.403	0.957	0.978
CoT + SFT	0.670	0.598	0.535	0.420	0.772	0.741
	0.855	0.842	0.593	0.429	0.960	0.980
PPO + BT reward	0.478	0.319	0.618	0.519	0.700	0.693
	0.483	0.381	0.705	0.468	0.933	0.966
RLAIF (CNF rewards)	0.475	0.340	0.517	0.349	0.773	0.762
	0.490	0.401	0.625	0.390	0.900	0.947
TRACE (Ours)	**0.755**	**0.688**	**0.842**	**0.653**	**0.807**	**0.771**
	0.935	**0.926**	**0.892**	**0.662**	**0.975**	**0.987**

Table 2: Ablation accuracy and macro-F1 (RL row above behavior row).

Method	RRT		NIV		OVISS	
	Acc.	F1	Acc.	F1	Acc.	F1
w/o PPO	0.742	0.672	0.705	0.557	0.800	0.780
	0.897	0.908	0.788	0.603	0.917	0.957
w/o concepts	0.612	0.562	0.532	0.409	0.818	0.795
	0.838	0.830	0.637	0.563	0.983	0.992
w/o behavior policy	0.773	0.713	0.852	0.707	0.820	0.805
	–	–	–	–	–	–

We evaluate explanation faithfulness by asking an independent judge to recover the target policy's action from the state and generated explanation. Recovery performance is evaluated by **Accuracy (Acc.)** and **macro-F1**, for the RL and behavior policies, respectively.

Table 1 compares TRACE with the baseline methods across the three clinical cohorts. TRACE achieves the highest action recovery accuracy and macro-F1 for both learned and behavior policies. For the learned policy, TRACE improves accuracy over the strongest baseline by 5.8, 21.0, and 3.5 percentage points on RRT, NIV, and OVISS, with corresponding macro-F1 gains of 0.049, 0.164, and 0.030. Similar improvements are observed for the behavior policy. On RRT, TRACE also achieves the highest F1 across all three actions, including an IHD recall of 0.949 (Supplementary Table S3), indicating that the gains are not driven solely by the dominant action.

The comparison also reveals that task specific supervision alone is not sufficient to match TRACE. SFT substantially improves over zero-shot prompting, whereas CoT+SFT provides no consistent additional benefit. Interpretable surrogate methods, including BC-LLM and Policy Distillation + LLM, remain below TRACE, suggesting that compressing a policy into an interpretable intermediate representation does not necessarily preserve enough information for faithful natural language action recovery. Reward optimized baselines also remain weaker: PPO with a Bradley–Terry reward increasingly produces generic monitoring language on RRT, while the RLAIF (CNF reward) is weakly sensitive to policy specific evidence and can amplify reader bias during optimization.

Ablations. To isolate the contribution of each major component of TRACE, we conduct three ablations. (1) **w/o PPO** removes PPO and directly evaluates the concept-conditioned SFT explainer. (2) **w/o concepts** removes the concept input, leaving the explainer to rely only on the state information. (3) **w/o behavior policy** applies the entire pipeline only to the learned policy, testing the contribution of joint training with behavior-policy explanations.

As shown in Table 2, removing concepts reduces learned-policy accuracy on RRT and NIV, while removing PPO also produces clear declines on these two cohorts. The effects are smaller on OVISS, consistent with its binary and highly concentrated action distribution, where only 2.7% of states initiate vasopressin (Supplementary Table S7). Training only on the learned policy remains competitive, indicating that joint training with the behavior policy is not required for strong RL action recovery. However, explaining the learned-policy jointly with the behavior-policy is meaningful in practice to allow human experts to accept or reject a different action proposed by the learned-policy, by understanding their difference in natural language.

3.2 Decision Relevance of Cited Concepts

To test whether the concepts emphasized in an explanation are functionally involved in the decision process, we conduct interventions on both the policy and the explainer. On the policy side, we remove the cited SAE features from the frozen representation, $h - \sum_{k \in S} z_k d_k$,where z_k is the activation of concept k and d_k is its decoder dictionary vector, and then re-run the complete policy decision rule. We measure the resulting *action flip rate*, i.e., the fraction of states for which the RL action changes after removal (Figure 2a). Across the 119 RRT disagreement states, removing the two or three cited concepts changes the RL action in 61.3% of states, compared with 40.2% for the same number of supplied but uncited concepts and 28.4% for random active concepts. This effect is nearly as large as removing all supplied concepts (63.0%) and is substantially larger than removing the same number of top concepts, indicating that the concepts emphasized by the explanation capture a concentrated set of decision-relevant features.

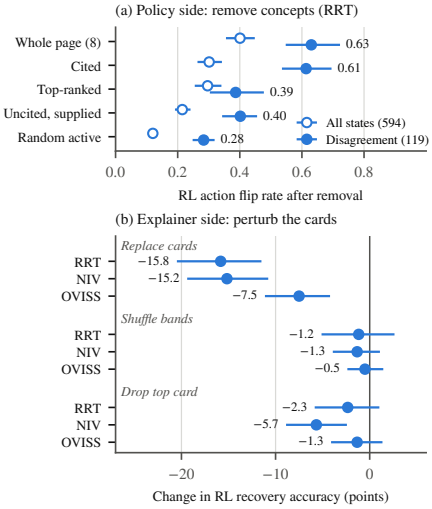


Figure 2: Decision relevance of the cited concepts. **(a)** Effect of removing cited concepts on RRT RL decisions. **(b)** Effect of perturbing concept cards on RL action recovery. Patient-clustered 95% intervals.

We next test whether the explainer itself depends on the concept evidence it receives, rather than producing plausible text from the state alone. We perturb the concept cards, regenerate the explanations, and measure the change in *RL action-recovery accuracy*, i.e., how often the judge can still recover the original RL action from the regenerated explanation (Figure 2b). Replacing the policy-selected cards with equally active but unselected concepts reduces RL recovery by 15.8, 15.2, and 7.5 percentage points on RRT, NIV, and OVISS, respectively. In contrast, removing only the top-ranked card produces smaller drops of 2.3, 5.7, and 1.3 points (Supplementary Tables S13 and S14). These results show that the explainer depends primarily on which concepts are provided, rather than only on their activation strength, and that the selected concept set carries information needed to preserve the target policy decision.

3.3 Supporting Clinical Reconsideration

We examine whether TRACE can support clinical reconsideration when the learned policy recommends a different action from the behavior policy. Given the learned policy's generally higher estimated value, our objective is to assess whether presenting explanations of both policies makes a convincing case for considering the learned policy's recommendation.

The two policies disagree on 19.0%, 11.9%, and 33.4% of decisions in the RRT, NIV, and OVISS test splits, respectively (Supplementary Table S6). In these cases, TRACE provides two explanations grounded in the same patient state. A clinician could use these paired explanations to compare the evidence supporting each choice and assess whether to follow the learned policy's alternative recommendation than the usual clinical practice.

We evaluate this by using LLM judges as a proxy for clinician assessment. Each judge sees the patient state and both explanations, blinded to their policy source, and then assigns a probability that one explanation makes more sense in terms of clinical coherence and grounding. On the 119 RRT disagreement states, Claude Opus 5 and GPT-6 Astra favor the learned-policy explanation on 74 and 70 states and the behavior-policy explanation on 32 and 49 states, respectively, a modest preference

under both judges (Figure 3). On the 76 NIV and 206 OVISS disagreement states, Claude Opus 5 favors the learned-policy explanation on 30 and 127 states and the behavior-policy explanation on 38 and 50 states, indicating no preference on NIV and a modest preference for the learned-policy explanation on OVISS (Figure S1). These comparisons suggest that TRACE can provide plausible reasons to reconsider some behavior-policy choices.

Table 3: Paired TRACE explanations for one RRT patient, excerpted where they differ. **Bold**: the measurements only that explanation cites, and the management it supports.

Patient	Age 61; 6-hour urine output 0 mL; creatinine 5.6 mg/dL; BUN 62 mg/dL; pH 7.29; bicarbonate 19 mEq/L; K 4.3 mEq/L; P/F ratio 112.5; BP 134.5/70 mmHg (mean 88).
No RRT *behavior policy*	… The cohort-relative renal solutes, urine output, electrolytes, and blood pressure pattern links creatinine 5.6 mg/dL, **potassium 4.3 mEq/L, diastolic pressure 70 mmHg**, and 0 mL/6h urine output, … and blood pressure pattern retains **systolic pressure 134.5 mmHg**. These measurements support **serial observation and reassessment** of the immediate corrective burden.
IHD *learned policy*	… The cohort-relative renal solutes, urine output, and oxygenation pattern emphasizes higher creatinine with lower urine output and **P/F ratio** … This joint representation supports the usefulness and tolerability of **efficient short-interval correction**.

A practical example. Table 3 illustrates how TRACE could support reconsideration when the behavior policy selects No RRT and the learned policy selects IHD for the same patient. Both explanations cite a concept linking renal solutes, acid-base status, and blood pressure, while the learned-policy explanation additionally cites a concept linking high creatinine with absent urine output and poor oxygenation that is not among the behavior policy's selected concepts. By showing how the policies group the same measurements differently, TRACE gives the clinician concrete reasons to assess when considering IHD rather than continued observation.

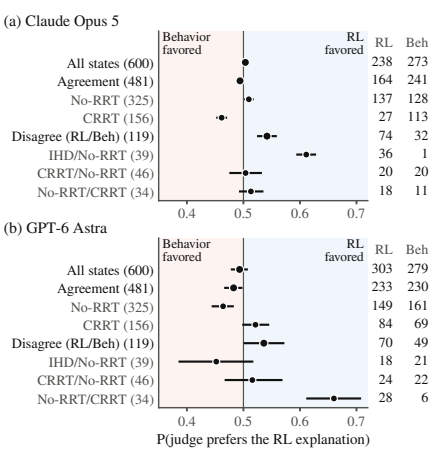


Figure 3: Same 600 RRT test states, judged blind to source in both orders by **(a)** Claude Opus 5 and **(b)** GPT-6 Astra: mean probability that the RL explanation is more sensible; states favoring each side at right. 95% patient-cluster intervals.

4 Conclusion

We introduced TRACE, a framework for explaining and comparing learned RL and behavior policies through natural language grounded in a shared concept space. Across three offline clinical RL benchmarks, TRACE achieves stronger action recovery than the baselines. Concept interventions demonstrate that the selected concepts are functionally relevant to both policy decisions and explanation recovery. Paired explanation comparisons further reveal how the two policies differ on the same state and where those differences are most apparent. Together, TRACE provides a practical framework for understanding what evidence each policy relies on and for auditing where and why learned policies diverge from observed behavior.

Our current pipeline still has several limitations. First, we currently work on discrete action spaces. Extending the framework to continuous action spaces remains more challenging, particularly when distinguishing decisions near action boundaries. Second, our current concepts are extracted from a fixed policy after training and therefore remain static. As a result, they do not capture how concepts evolve as the policy changes during RL training. We will try to study concept dynamics across training checkpoints to better understand how policy representations develop over time in the future.

ETHICS STATEMENT

This work is a retrospective computational study using deidentified ICU records from MIMIC-III and MIMIC-IV (Johnson et al., 2016a; 2024). These databases are available through credentialed access on PhysioNet; access, use, and sharing of patient-derived materials are subject to the applicable data use agreements. The study evaluates explanations of frozen policies offline and does not involve prospective patient intervention. The policies and generated explanations are research artifacts, not validated recommendations for patient treatment. Historical treatment patterns can reflect confounding, unequal access to care, and demographic biases, while fluent explanations can encourage unwarranted trust in an incorrect decision. Action recovery therefore does not establish clinical benefit, safety, or fairness. Any clinical application would require independent clinical review, subgroup evaluation, prospective validation, and appropriate privacy and institutional oversight.

REPRODUCIBILITY STATEMENT

Anonymized source code for TRACE is provided in the supplementary material. For all three cohorts it implements every stage of Algorithm 1: training of the shared-encoder policy with its learned-policy and behavior branches, fitting and selection of the policy-preserving SAE, concept-dictionary construction and policy-specific concept selection, teacher supervision with its acceptance checks, explainer SFT and shared PPO with the reader reward, explanation generation, and the decision-effect computation used for policy comparison. The code contains all prompt templates of the method (teacher, reader, and concept naming), the cohort-construction code for RRT and NIV, and one script per pipeline stage with the hyperparameters reported here. Upon publication, we will also release the judging code, the vasopressin cohort extraction, the three concept dictionaries with their rendered cards, and the frozen policy, SAE, and explainer checkpoints (SFT and PPO), as permitted by the MIMIC data use agreements. No patient-level data are distributed; reproducing data-dependent experiments requires credentialed access to the underlying MIMIC databases.

GENERATIVE AI USE STATEMENT

Generative AI tools were used solely to assist with language editing and polishing of this manuscript, including improvements to grammar, clarity, and readability. They were not used to generate the research ideas, formulate the methodology, conduct the experiments, analyze the data, or determine the scientific conclusions. All AI-assisted suggestions were reviewed and revised by the authors, who take full responsibility for the content of the paper.

REFERENCES

Sahar Admoni, Assaf Hallak, Yftah Ziser, Omer Ben-Porat, and Ofra Amir. From actions to words: Towards abstractive-textual policy summarization in RL. In *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems*, 2026. doi: 10.65109/EGVR4787.

Yotam Amitai and Ofra Amir. "I don't think so": Summarizing policy disagreements for agent comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 5269–5276, 2022. doi: 10.1609/aaai.v36i5.20463.

Yotam Amitai, Yael Septon, and Ofra Amir. Explaining reinforcement learning agents through counterfactual action outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 10003–10011, 2024. doi: 10.1609/aaai.v38i9.28863.

Anthropic. Introducing Claude Opus 5. `https://www.anthropic.com/news/claude-opus-5`, July 2026a. Accessed: 2026-09-19.

Anthropic. Introducing Claude Sonnet 5. `https://www.anthropic.com/news/claude-sonnet-5`, June 2026b. Accessed: 2026-09-25.

Sanjay Basu, Sadiq Patel, Parth Sheth, Bhairavi Muralidharan, Namrata Elamaran, Aakriti Kinra, and Rajaie Batniji. Mechanistic interpretability of reinforcement learning in Medicaid care coordination. *BMJ Health & Care Informatics*, 33(1):e101935, 2026. doi: 10.1136/bmjhci-2025-101935.

Jenny M. Bauer and Martin Michalowski. Human-centered explainability evaluation in clinical decision-making: a critical review of the literature. *Journal of the American Medical Informatics Association*, 32(9):1477–1484, 2025. doi: 10.1093/jamia/ocaf110.

Yanzhe Bekkemoen. Explainable reinforcement learning (XRL): a systematic literature review and taxonomy. *Machine Learning*, 113(1):355–441, 2024. doi: 10.1007/s10994-023-06479-7.

Ayoub Belouadah, Marcelo Luis Ruiz-Rodríguez, Sylvain Kubler, and Yves Le Traon. Evaluating the effectiveness of LLMs for explainable deep reinforcement learning. *Machine Learning with Applications*, 22:100795, 2025. doi: 10.1016/j.mlwa.2025.100795.

Trenton Bricken, Adly Templeton, Joshua Batson, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL `https://transformer-circuits.pub/2023/monosemantic-features/index.html`.

Quentin Delfosse, Sebastian Sztwiertnia, Mark Rothermel, Wolfgang Stammer, and Kristian Kersting. Interpretable concept bottlenecks to align reinforcement learning agents. In *Advances in Neural Information Processing Systems*, volume 37, pp. 66826–66855, 2024. doi: 10.52202/079017-2135.

Upol Ehsan, Brent Harrison, Larry Chan, and Mark O. Riedl. Rationalization: A neural machine translation approach to generating natural language explanations. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 81–87, 2018. doi: 10.1145/3278721.3278736.

Jean Feng, Avni Kothari, Lucas Zier, Chandan Singh, and Yan Shuo Tan. Bayesian concept bottleneck models with LLM priors. In *Advances in Neural Information Processing Systems*, volume 38, pp. 93889–93920, 2025.

Xue Feng, Daoyuan Wang, Qing Pan, Molei Yan, Xiaoqing Liu, Yanfei Shen, Luping Fang, Guolong Cai, and Gangmin Ning. Reinforcement learning model for managing noninvasive ventilation switching policy. *IEEE Journal of Biomedical and Health Informatics*, 27(8):4120–4130, 2023. doi: 10.1109/JBHI.2023.3274568.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International conference on machine learning*, pp. 2052–2062. PMLR, 2019.

Jasmina Gajcin, Rahul Nair, Tejaswini Pedapati, Radu Marinescu, Elizabeth Daly, and Ivana Dusparic. Contrastive explanations for comparing preferences of reinforcement learning agents. *arXiv preprint arXiv:2112.09462*, 2021.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, 2025.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23), 2000. doi: 10.1161/01.cir.101.23.e215.

Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nature Medicine*, 25(1):16–18, 2019. doi: 10.1038/s41591-018-0310-5.

Bálint Gyevnár, Christopher G. Lucas, Stefano V. Albrecht, and Shay B. Cohen. Integrating counterfactual simulations with language models for explaining multi-agent behaviour. In *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems*, 2026. doi: 10.65109/MCSD1905.

Peter Hase and Mohit Bansal. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5540–5552. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.491.

Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4351–4367. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.findings-emnlp.390.

Bradley Hayes and Julie A. Shah. Improving robot controller transparency through autonomous policy explanation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 303–312, 2017. doi: 10.1145/2909824.3020233.

William Hua, Hongyuan Mei, Sarah Zohar, Magali Giral, and Yanxun Xu. Personalized dynamic treatment regimes in continuous time: a Bayesian approach for optimizing clinical decisions with timing. *Bayesian Analysis*, 17(3):849–878, 2022. doi: 10.1214/21-BA1276.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, 2024.

David R. Hunter. MM algorithms for generalized Bradley-Terry models. *The Annals of Statistics*, 32(1):384–406, 2004. doi: 10.1214/aos/1079120141.

Pushkala Jayaraman, Jacob Desman, Moein Sabounchi, Girish N Nadkarni, and Ankit Sakhuja. A primer on reinforcement learning in medicine for clinicians. *npj Digital Medicine*, 7(1):337, 2024. doi: 10.1038/s41746-024-01316-0.

Zhengping Jiang, Yining Lu, Hanjie Chen, Daniel Khashabi, Benjamin Van Durme, and Anqi Liu. RORA: Robust free-text rationale evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1070–1087. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.60.

Alistair Johnson, Tom Pollard, and Roger Mark. MIMIC-III Clinical Database. *PhysioNet*, September 2016a. doi: 10.13026/C2XW26. URL https://doi.org/10.13026/C2XW26. Version 1.4.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV. *PhysioNet*, October 2024. doi: 10.13026/kpb9-mt58. URL https://doi.org/10.13026/kpb9-mt58. Version 3.1.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1), 2023. doi: 10.1038/s41597-022-01899-x.

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1), 2016b. doi: 10.1038/sdata.2016.35.

Alexandre Kalimouttou, Jason N. Kennedy, Jean Feng, Harvineet Singh, Suchi Saria, Derek C. Angus, Christopher W. Seymour, and Romain Pirracchio. Optimal vasopressin initiation in septic shock: The OVISS reinforcement learning study. *JAMA*, 333(19):1688–1698, 2025. doi: 10.1001/jama.2025.3046.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2668–2677. PMLR, 2018.

Haechang Kim, Hao Chen, Can Li, and Jong Min Lee. TalkToAgent: A multi-agent LLM framework for natural language explanation of reinforcement learning policies. *Computers & Chemical Engineering*, 211:109672, 2026. doi: 10.1016/j.compchemeng.2026.109672.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 57(9):242:1–242:52, 2025. doi: 10.1145/3728458.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348. PMLR, 2020.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pp. 22199–22213, 2022.

Matthieu Komorowski, Leo A. Celi, Omar Badawi, Anthony C. Gordon, and A. Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716–1720, 2018. doi: 10.1038/s41591-018-0213-5.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Lada Lijović, Tomislav Radočaj, Nataša Kovač, Marinko Vučić, and Paul Elbers. Predictive performance of ROX index and its variations for NIV failure. *Medicina Intensiva (English Edition)*, 49 (7):502136, 2025. doi: 10.1016/j.medine.2025.502136.

Lirui Luo, Guoxi Zhang, Hongming Xu, Yaodong Yang, Cong Fang, and Qing Li. End-to-end neuro-symbolic reinforcement learning with textual explanations. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 33533–33557. PMLR, 2024a. URL https://proceedings.mlr.press/v235/luo24j.html.

Zhiyao Luo, Yangchen Pan, Peter Watkinson, and Tingting Zhu. Position: Reinforcement learning in dynamic treatment regimes needs critical reexamination. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 33432–33465. PMLR, 2024b. URL https://proceedings.mlr.press/v235/luo24f.html.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. Towards faithful model explanation in NLP: A survey. *Computational Linguistics*, 50(2):657–723, 2024. doi: 10.1162/coli_a_00511.

Alireza Makhzani and Brendan Frey. k-sparse autoencoders. In *2nd International Conference on Learning Representations (ICLR)*, 2014. URL https://arxiv.org/abs/1312.5663.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *International Conference on Learning Representations*, 2025.

Katie Matton, Robert Ness, John Guttag, and Emre Kiciman. Walk the talk? Measuring the faithfulness of large language model explanations. In *International Conference on Learning Representations*, pp. 73212–73277, 2025. URL https://openreview.net/forum?id=4ub9gpx9xw.

Daniele Melloni and Andrea Zingoni. A survey on explainable reinforcement learning: state of the art, challenges and opportunities. *Data Mining and Knowledge Discovery*, 40(5):76, 2026. doi: 10.1007/s10618-026-01246-3.

Stephanie Milani, Nicholay Topin, Manuela Veloso, and Fei Fang. Explainable reinforcement learning: A survey and comparative review. *ACM Computing Surveys*, 56(7):1–36, 2024. doi: 10.1145/3616864.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019. doi: 10.1016/j.artint.2018.07.007.

Julian Minder, Clément Dumas, Caden Juang, Bilal Chughtai, and Neel Nanda. Overcoming sparsity artifacts in crosscoders to interpret chat-tuning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 117842–117893, 2025. doi: 10.52202/085713-3551.

S. A. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 65(2):331–355, 2003. doi: 10.1111/1467-9868.00389.

Hongseok Namkoong, Ramtin Keramati, Steve Yadlowsky, and Emma Brunskill. Off-policy policy evaluation for sequential decisions under unobserved confounding. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *International Conference on Learning Representations*, 2023.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. https://openai.com/index/gpt-5-6/, July 2026a. Accessed: 2026-09-19.

OpenAI. GPT-6 Astra: A new generation of intelligence. https://openai.com/index/gpt-6-astra/, September 2026b. Accessed: 2026-09-25.

Martijn Otten, Ameet R. Jagesar, Tariq A. Dam, Laurens A. Biesheuvel, Floris den Hengst, Kirsten A. Ziesemer, Patrick J. Thoral, Harm-Jan de Grooth, Armand R. J. Girbes, Vincent François-Lavet, Mark Hoogendoorn, and Paul W. G. Elbers. Does reinforcement learning improve outcomes for critically ill patients? A systematic review and level-of-readiness assessment. *Critical Care Medicine*, 52(2):e79–e88, 2024. doi: 10.1097/CCM.0000000000006100.

Letitia Parcalabescu and Anette Frank. On measuring faithfulness or self-consistency of natural language explanations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6048–6089. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.329.

Arne Peine, Ahmed Hallawa, Johannes Bickenbach, Guido Dartmann, Lejla Begic Fazlic, Anke Schmeink, Gerd Ascheid, Christoph Thiemermann, Andreas Schuppert, Ryan Kindle, Leo Celi, Gernot Marx, and Lukas Martin. Development and validation of a reinforcement learning algorithm to dynamically optimize mechanical ventilation in critical care. *npj Digital Medicine*, 4(1): 32, 2021. doi: 10.1038/s41746-021-00388-6.

Eleonora Poeta, Gabriele Ciravegna, Eliana Pastor, Tania Cerquitelli, and Elena Baralis. Concept-based explainable artificial intelligence: A survey. *ACM Computing Surveys*, 2025. doi: 10.1145/3774643.

Hangqi Ren and Junyi Liao. Unmasking toxic mimicry in medical offline reinforcement learning for ICU sepsis management via counterfactual clinical audits. *arXiv preprint arXiv:2608.11410*, 2026.

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

Lisa Schut, Nenad Tomašev, Thomas McGrath, Demis Hassabis, Ulrich Paquet, and Been Kim. Bridging the human–AI knowledge gap through concept discovery and transfer in AlphaZero. *Proceedings of the National Academy of Sciences*, 122(13):e2406675122, 2025. doi: 10.1073/pnas.2406675122.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. In *Advances in Neural Information Processing Systems*, volume 35, pp. 9460–9471, 2022. doi: 10.52202/068431-0687.

Ziyi Tang, Peilin Sun, Ruitong Liang, Siqi Tong, Yi Yang, Meicheng Yang, and Hui Chen. Reinforcement learning for treatment decision-making in sepsis: a scoping review. *npj Digital Medicine*, 2026. doi: 10.1038/s41746-026-03034-1.

Harrish Thasarathan, Julian Forsyth, Thomas Fel, Matthew Kowal, and Konstantinos G. Derpanis. Universal sparse autoencoders: Interpretable cross-model concept alignment. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 59304–59325. PMLR, 2025. URL https://proceedings.mlr.press/v267/thasarathan25a.html.

Sana Tonekaboni, Shalmali Joshi, Melissa D. McCradden, and Anna Goldenberg. What clinicians want: Contextualizing explainable machine learning for clinical end use. In *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106 of *Proceedings of Machine Learning Research*, pp. 359–380. PMLR, 2019.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, pp. 74952–74965, 2023. doi: 10.52202/075280-3275.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Qianyi Xu, Feng Wu, Zi Yi Christopher Thong, Mark Sen Liang Goh, Pengpeng Chen, Jie Yang, Chen Huang, Zhongheng Zhang, Yucai Hong, Kay Choong See, and Mengling Feng. HRRT: Hierarchical reinforcement learning for renal replacement therapy decision support. *npj Digital Medicine*, 2026. doi: 10.1038/s41746-026-02900-2.

Xinyi Yang, Liang Zeng, Heng Dong, Chao Yu, Xiaoran Wu, Huazhong Yang, Yu Wang, Milind Tambe, and Tonghan Wang. Translate policy to language: Flow matching generated rewards for LLM explanations. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=zmZsWCGzUV.

Sophia Yazzourh, Nicolas Savy, Philippe Saint-Pierre, and Michael R. Kosorok. Medical knowledge integration into reinforcement learning algorithms for dynamic treatment regimes. *International Statistical Review*, 94(2):382–411, 2026. doi: 10.1111/insr.12617.

Muhammad Hamza Yousuf, Jason Li, Sahar Vahdati, Raphael Theilen, Jakob Wittenstein, and Jens Lehmann. Advancing safe mechanical ventilation using offline RL with hybrid actions and clinically aligned rewards. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 39548–39557, 2026. doi: 10.1609/aaai.v40i46.41306.

Xi-Jia Zhang, Yue Guo, Shufei Chen, Simon Stepputtis, Matthew Gombolay, Katia Sycara, and Joseph Campbell. Model-agnostic policy explanations with large language models. In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=VzXpFjKgJg.

Supplementary Material

TABLES AND FIGURES CITED IN THE MAIN PAPER

The tables and figures that the main paper cites are collected here, in the order in which the main paper first cites them. The sections that follow describe how each was produced and discuss it further.

Table S1: Held-out SAE fidelity per cohort. W/K is dictionary width / active coordinates per state; agreement is the fraction of test states on which the frozen branch, re-run on the reconstructed representation, selects the same action as on the original one; centered-Q corr. is the correlation between original and reconstructed action-centered Q values. Kendall τ is the per-state rank correlation between the actions ordered by the original and by the reconstructed Q values, averaged over test states.

Cohort	W/K	Hidden R^2	RL agreement	Behavior agreement	Centered-Q corr.	Kendall τ
RRT	512 / 128	0.998	0.991	0.996	1.000	0.982
NIV	512 / 128	0.995	0.995	0.997	1.000	0.998
Vasopressin	64 / 6	0.984	0.982	0.999	0.995	0.972

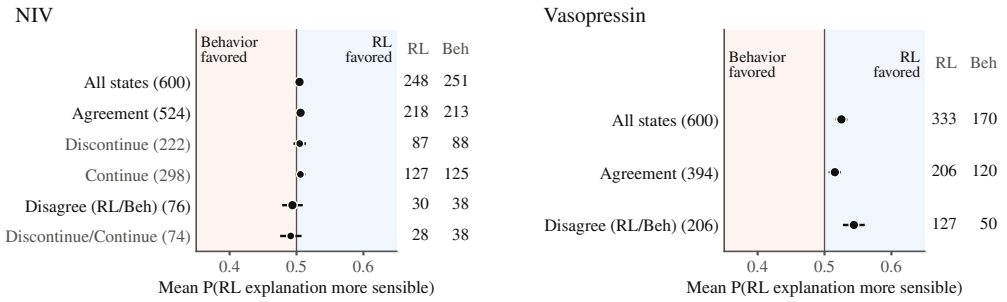


Figure S1: Blinded head-to-head explanation preference on the NIV and vasopressin test cohorts (600 states each; the RRT preference is in main Figure 3): per-state probability that the RL explanation makes more sense, averaged over both orders; the columns count states favoring each side. Patient-cluster 95% intervals.

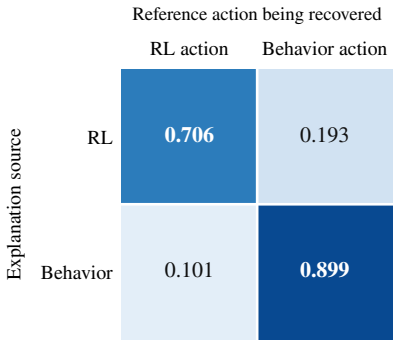


Figure S2: Cross-policy action recovery on RRT disagreement states. Diagonal entries measure recovery from policy-matched explanations, while off-diagonal entries measure recovery from explanations of the alternative policy.

Table S2: One training state of the RRT teacher data (training split) on which the two policies disagree. Each card names the clinical axes its SAE coordinate moves (every name reads "Cohort-relative … pattern") and states the patient's activation band with the direction that band carries relative to the naming cohort. Sex and race codes, also in the state, are omitted here.

Instruction	Explain how the frozen renal-support policy weighs the patient-specific evidence without revealing its selected action.	
Patient state	Age 73 years; weight 73.5 kg; heart rate 69.5 bpm; pH 7.14; blood pressure 52.667/30.5 mmHg (mean 38.333); SpO_2 99%; BUN 27 mg/dL; creatinine 2.1 mg/dL; bicarbonate 9 mEq/L; potassium 5.1 mEq/L; sodium 133.0 mEq/L; P/F ratio 330.0; 6-hour urine output 0 mL/6h; weight-normalized 6-hour urine output 0 mL/kg/h.	
	Learned (RL) policy	**Behavior policy**
Policy line	Target policy: RL — the frozen reinforcement-learning treatment policy.	Target policy: Behavior — the frozen clinician-behavior model.
Concept cards	1. acid-base status, blood pressure, and oxygenation | HIGH: lower bicarbonate, systolic blood pressure, and SpO_2 2. urine output, renal solutes, blood pressure, and electrolytes | HIGH: lower 6-hour urine output, BUN, systolic blood pressure, and sodium **3.** acid-base status, blood pressure, and electrolytes | HIGH: higher potassium; lower bicarbonate and diastolic blood pressure | bicarbonate: 9 mEq/L; diastolic blood pressure: 30.5 mmHg; potassium: 5.1 mEq/L **4.** blood pressure, acid-base status, urine output, and oxygenation | HIGH: lower systolic blood pressure, bicarbonate, 6-hour urine output, and P/F ratio | systolic blood pressure: 52.667 mmHg; bicarbonate: 9 mEq/L; 6-hour urine output: 0 mL/6h; P/F ratio: 330.0 **5.** acid-base status, renal solutes, and blood pressure | HIGH: higher BUN; lower bicarbonate and mean blood pressure | bicarbonate: 9 mEq/L; BUN: 27 mg/dL; mean blood pressure: 38.333 mmHg 6. renal solutes, blood pressure, and oxygenation | HIGH: higher SpO_2; lower BUN and diastolic blood pressure 7. renal solutes, heart rate, and urine output | HIGH: higher BUN, heart rate, and 6-hour urine output 8. blood pressure, urine output, renal solutes, and acid-base status | HIGH: lower systolic blood pressure, 6-hour urine output, BUN, and bicarbonate	**1.** acid-base status, blood pressure, and oxygenation | HIGH: lower bicarbonate, systolic blood pressure, and SpO_2 | bicarbonate: 9 mEq/L; systolic blood pressure: 52.667 mmHg; SpO_2: 99 % 2. urine output and electrolytes | HIGH: higher 6-hour urine output and sodium **3.** acid-base status, renal solutes, and blood pressure | HIGH: higher BUN; lower bicarbonate and mean blood pressure | bicarbonate: 9 mEq/L; BUN: 27 mg/dL; mean blood pressure: 38.333 mmHg 4. oxygenation, blood pressure, acid-base status, and electrolytes | HIGH: higher potassium; lower SpO_2, systolic blood pressure, and bicarbonate **5.** blood pressure, urine output, renal solutes, and acid-base status | HIGH: lower systolic blood pressure, 6-hour urine output, BUN, and bicarbonate | systolic blood pressure: 52.667 mmHg; 6-hour urine output: 0 mL/6h; BUN: 27 mg/dL; bicarbonate: 9 mEq/L **6.** blood pressure, renal solutes, urine output, and acid-base status | HIGH: higher creatinine; lower mean blood pressure, weight-normalized 6-hour urine output, and bicarbonate | mean blood pressure: 38.333 mmHg; creatinine: 2.1 mg/dL; weight-normalized 6-hour urine output: 0 mL/kg/h; bicarbonate: 9 mEq/L 7. blood pressure, renal solutes, and acid-base status | HIGH: lower systolic blood pressure, BUN, and bicarbonate 8. *Uncertain SAE feature (no name; not citable)*
Private target	CRRT	No RRT
Teacher explanation (SFT target)	The selected representation patterns are cohort-relative, emphasizing the joint acid-base, blood-pressure, oxygenation, urine-output, renal-solute, and electrolyte coordinates rather than clinical severity labels. The Cohort-relative acid-base status, blood pressure, and electrolytes pattern aligns higher potassium with lower bicarbonate and diastolic pressure, reflected by potassium 5.1 mEq/L, bicarbonate 9 mEq/L, and diastolic pressure 30.5 mmHg. The Cohort-relative blood pressure, acid-base status, urine output, and oxygenation pattern and Cohort-relative acid-base status, renal solutes, and blood pressure pattern additionally join systolic pressure 52.667 mmHg, zero 6-hour urine output, bicarbonate 9 mEq/L, and mean pressure 38.333 mmHg; together these support gentle, sustained rather than large short-interval physiologic shifts.	The selected representation patterns are cohort-relative and emphasize a convergent pattern spanning bicarbonate, systolic and mean pressure, urine output, creatinine, BUN, potassium, and oxygenation. They connect bicarbonate of 9 mEq/L, systolic pressure of 52.667 mmHg, mean pressure of 38.333 mmHg, and 0 mL/6h urine output with creatinine of 2.1 mg/dL and potassium of 5.1 mEq/L. The directions refer only to naming-cohort coordinates, so this convergence supports observing the immediate physiologic burden and reassessing its measured trajectory.

Table S3: Per-action recovery of the learned RRT policy on 600 test states. Frozen explainers appear above the rule and fine-tuned explainers below it.

Method	Acc.	No RRT			IHD			CRRT		
		Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Direct LLM	0.437	0.578	0.554	0.566	0.200	0.026	0.045	0.247	0.307	0.274
Zero-shot CoT	0.438	0.592	0.513	0.549	0.000	0.000	0.000	0.276	0.391	0.324
MCTS Refinement	0.373	0.531	0.457	0.491	0.067	0.026	0.037	0.214	0.292	0.247
BC-LLM	0.487	0.592	0.688	0.637	0.127	0.205	0.157	0.308	0.183	0.230
Policy Distillation + LLM	0.412	0.557	0.529	0.543	**0.333**	0.077	0.125	0.216	0.267	0.239
SFT	0.697	0.862	0.607	0.712	0.288	0.872	0.433	0.725	**0.822**	0.770
CoT + SFT	0.670	0.832	0.593	0.693	0.231	0.692	0.346	0.714	0.802	0.755
PPO + BT reward	0.478	0.586	0.685	0.632	0.273	0.077	0.120	0.225	0.188	0.205
RLAIF (CNF rewards)	0.475	0.656	0.462	0.542	0.125	0.026	0.043	0.348	0.584	0.436
TRACE (Ours)	**0.755**	**0.871**	**0.733**	**0.796**	0.298	**0.949**	**0.454**	**0.879**	0.757	**0.814**

Table S4: Per-action recovery of the learned NIV policy on 600 test states. Frozen explainers appear above the rule and fine-tuned explainers below it; the two ablation rows at the bottom remove one TRACE component each.

Method	Acc.	Discontinue			Continue			Intubate		
		Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Direct LLM	0.500	0.667	0.007	0.013	0.498	**0.993**	0.664	**0.667**	0.333	**0.444**
Zero-shot CoT	0.512	0.929	0.044	0.084	0.506	0.980	0.667	0.222	0.333	0.267
MCTS Refinement	0.528	0.719	0.139	0.232	0.518	0.919	0.663	0.143	0.333	0.200
BC-LLM	0.502	0.553	0.476	0.512	0.506	0.527	0.516	0.086	0.500	0.146
Policy Distillation + LLM	0.537	0.750	0.132	0.224	0.518	0.946	0.670	0.250	0.167	0.200
SFT	0.545	0.618	0.470	0.534	0.544	0.621	0.580	0.086	0.500	0.146
CoT + SFT	0.535	0.603	0.486	0.538	0.546	0.581	0.563	0.091	**0.667**	0.160
PPO + BT reward	0.618	0.724	0.443	0.549	0.589	0.792	0.675	0.222	**0.667**	0.333
RLAIF (CNF rewards)	0.517	**1.000**	0.047	0.090	0.509	0.987	0.671	0.250	0.333	0.286
TRACE (Ours)	**0.842**	0.932	**0.784**	**0.851**	**0.830**	0.903	**0.865**	0.148	**0.667**	0.242
TRACE w/o PPO (Concept SFT)	0.705	0.881	0.527	0.660	0.650	0.886	0.750	0.176	0.500	0.261
TRACE w/o concepts (state-only PPO)	0.532	0.654	0.338	0.445	0.530	0.721	0.611	0.098	0.667	0.170

Table S5: Per-action recovery of the learned vasopressin policy on 600 test states. Frozen explainers appear above the rule and fine-tuned explainers below it; the two ablation rows at the bottom remove one TRACE component each.

Method	Acc.	No vasopressin			Vasopressin		
		Precision	Recall	F1	Precision	Recall	F1
Direct LLM	0.678	0.734	0.799	0.765	0.538	0.447	0.488
Zero-shot CoT	0.720	0.746	0.871	0.803	0.636	0.432	0.514
MCTS Refinement	0.500	0.846	0.292	0.434	0.399	**0.898**	0.552
BC-LLM	0.617	0.828	0.525	0.643	0.466	0.791	0.586
Policy Distillation + LLM	0.683	0.827	0.655	0.731	0.528	0.738	0.615
SFT	0.765	0.799	0.858	0.827	0.684	0.587	0.632
CoT + SFT	0.772	0.813	0.848	0.830	0.683	0.626	0.653
PPO + BT reward	0.700	0.864	0.645	0.738	0.542	0.806	0.648
RLAIF (CNF rewards)	0.773	**0.882**	0.756	0.814	0.634	0.806	**0.709**
TRACE (Ours)	**0.807**	0.814	**0.914**	**0.861**	**0.785**	0.602	0.681
TRACE w/o PPO (Concept SFT)	0.800	0.853	0.840	0.847	0.703	0.723	0.713
TRACE w/o concepts (state-only PPO)	0.818	0.848	0.881	0.864	0.754	0.699	0.725

Table S6: Policy disagreement on each cohort's held-out test split (18,654 / 1,794 / 5,076 decisions), by direction (behavior action → learned action). Vasopressin counts decision states only.

Cohort	Disagree	Behavior → learned	n
RRT	19.0%	CRRT → No RRT	1,329
		No RRT → IHD	1,086
		No RRT → CRRT	1,129
		All disagreements	3,549
NIV	11.9%	Continue → Discontinue	208
		All disagreements	213
Vasopressin	33.4%	No vasopressin → Initiate	1,694

Table S7: Action distributions over all decision states of each cohort. Each block gives the recorded clinician action and the two frozen explanation targets. Cells are counts with the share of states in parentheses, rounded to three decimals. The vasopressin cohort has two actions, so its third column is unused. RRT and NIV counts pool the training, validation, and test splits. Vasopressin counts cover decision-eligible epochs only (57,574 of 155,820 extracted epochs); post-initiation epochs are retained for outcome follow-up but are not initiation opportunities.

Source	Action count (share of states)		
RRT modality (n = 114,113): No RRT / IHD / CRRT			
Recorded clinician	67,359 (0.590)	3,984 (0.035)	42,770 (0.375)
Estimated behavior	73,112 (0.640)	2,182 (0.020)	38,819 (0.340)
Frozen RL (BCQ)	69,348 (0.608)	5,723 (0.050)	39,042 (0.342)
NIV switching (n = 18,320): discontinue / continue NIV / intubate			
Recorded clinician	8,001 (0.437)	10,003 (0.546)	316 (0.017)
Estimated behavior	6,884 (0.376)	11,268 (0.615)	168 (0.009)
Frozen RL (BCQ)	9,022 (0.492)	9,107 (0.497)	191 (0.010)
Vasopressin initiation (n = 57,574): norepinephrine alone / initiate vasopressin			
Recorded clinician	56,024 (0.973)	1,550 (0.027)	
Estimated behavior	57,546 (>0.999)	28 (<0.001)	
Frozen RL (BCQ)	38,584 (0.670)	18,990 (0.330)	

Table S8: Paired matched-query outcomes on the same 119 RRT disagreement states. Each query uses its corresponding policy tag. Counts in each row sum to 119; both-correct recovery requires both policies' distinct actions to be recovered for the same state.

Checkpoint	Both correct	RL only	Behavior only	Neither	Both correct (%)
Concept SFT	53	16	41	9	44.5
TRACE (2604)	74	10	33	2	62.2

Table S9: Iteration-2604 action recovery by policy agreement. Cross follows other is the fraction of cross-arm predictions matching the other policy's action. Agreement and disagreement are defined by the two frozen policies before explanation generation.

Policy	Subset	States	State only	Matched	Cross	Cross follows other
RL	All	600	0.453	0.755	0.777	0.935
RL	Agreement	481	0.455	0.767	0.944	0.944
RL	Disagreement	119	0.445	0.706	0.101	0.899
Behavior	All	600	0.488	0.935	0.652	0.758
Behavior	Agreement	481	0.514	0.944	0.765	0.765
Behavior	Disagreement	119	0.387	0.899	0.193	0.731

Table S10: Policy specificity on the disagreement states of the 600-state test cohorts: recovery of each policy's action, with the judge told that policy, from its own explanation, from the other policy's explanation of the same state, and from the state alone (state and policy tag, no explanation). The RRT own and other cells are the heat map of Figure S2. On NIV every behavior disagreement target is Continue, which the state-only judge already answers, and on vasopressin the behavior target is a single class (Section S1); on these two the informative comparison is the drop from Own to Other.

	RL			Behavior		
Cohort (disagreement states)	Own	Other	State only	Own	Other	State only
RRT (119)	0.706	0.101	0.445	0.899	0.193	0.387
NIV (76)	0.513	0.000	0.211	1.000	0.474	1.000
Vasopressin (206)	0.602	0.150	0.534	0.961	0.432	0.903

Table S11: Intervention on the concepts an explanation *cites*. For every state of the 600-state test cohorts, the cited set is recovered from the branch's own TRACE explanation by the model-free matcher (strict rule, Table S23) and removed from the shared representation, $h_S = h - \sum_{k \in S} z_k d_k$, before the complete frozen decision rule is re-run. Every size-matched condition removes m concepts, m being the size of the cited set: the m top-ranked page cards, m uncited page cards, or m random active concepts (the last two average 20 draws per state); the *whole page* row removes all k cards. *Flip* is the rate at which the deployed action changes. States whose explanation cites nothing, or whose page holds fewer than m uncited cards, are excluded; n gives the states that remain. Brackets: patient-clustered bootstrap 95% intervals (2,000 draws). RL branch; the behavior-branch contrast on RRT is given in Section S3.2.

Cohort	Branch	Removed concepts	Flip (all)	Flip (disagree)	Δ margin (all)	n (all/dis.)
RRT ($k = 8$)	RL ($\bar{m} = 2.5$)	whole page (k)	0.401 [0.355, 0.448]	0.630 [0.548, 0.723]	1.12 [0.84, 1.40]	594/119
		cited (m)	0.301 [0.264, 0.342]	0.613 [0.536, 0.695]	0.37 [0.21, 0.54]	594/119
		top-m ranked page cards	0.296 [0.255, 0.341]	0.387 [0.304, 0.477]	0.65 [0.47, 0.84]	594/119
		uncited page cards (m)	0.215 [0.191, 0.241]	0.402 [0.343, 0.456]	0.25 [0.15, 0.35]	594/119
		random active (m)	0.120 [0.106, 0.134]	0.284 [0.248, 0.319]	0.02 [0.01, 0.04]	594/119
NIV ($k = 8$)	RL ($\bar{m} = 1.2$)	whole page (k)	0.083 [0.053, 0.116]	0.075 [0.000, 0.170]	0.11 [0.01, 0.23]	327/40
		cited (m)	0.015 [0.003, 0.033]	0.025 [0.000, 0.088]	0.00 [-0.03, 0.03]	327/40
		top-m ranked page cards	0.018 [0.006, 0.035]	0.025 [0.000, 0.088]	0.09 [0.04, 0.13]	327/40
		uncited page cards (m)	0.019 [0.010, 0.031]	0.046 [0.009, 0.105]	0.02 [0.01, 0.04]	327/40
		random active (m)	0.009 [0.004, 0.015]	0.021 [0.004, 0.047]	0.01 [0.00, 0.01]	327/40
Vasopressin ($k = 4$)	RL ($\bar{m} = 1.0$)	whole page (k)	0.840 [0.777, 0.895]	0.883 [0.805, 0.944]	0.40 [0.26, 0.56]	262/111
		cited (m)	0.534 [0.462, 0.604]	0.847 [0.769, 0.912]	-0.12 [-0.84, 0.60]	262/111
		top-m ranked page cards	0.431 [0.356, 0.511]	0.901 [0.837, 0.953]	-3.83 [-5.06, -2.37]	262/111
		uncited page cards (m)	0.421 [0.373, 0.468]	0.684 [0.634, 0.732]	0.08 [-0.20, 0.37]	262/111
		random active (m)	0.322 [0.294, 0.351]	0.506 [0.473, 0.538]	0.10 [-0.02, 0.20]	262/111

Table S12: Paired contrasts of the cited-concept intervention (RL branch), under both citation rules. Each cell is the per-state difference in action-flip rate between two size-matched removals, resampled by patient (2,000 draws). *cited − uncited page* is the most direct contrast: same page, same selector, different assertion. *cited − top-m ranked* asks whether citation carries anything the page's own ordering does not. *whole page − cited* is what whole-page removal adds over the asserted subset; a value near zero means the two tests agree.

		strict rule		coverage rule	
Cohort	Contrast	All states	Disagreement	All states	Disagreement
RRT	cited − uncited page	0.087 [0.042, 0.128]	0.212 [0.109, 0.319]	0.093 [0.049, 0.139]	0.290 [0.176, 0.407]
	cited − top-m ranked	0.005 [-0.041, 0.045]	0.227 [0.130, 0.321]	-0.007 [-0.052, 0.034]	0.307 [0.188, 0.426]
	cited − random active	0.181 [0.146, 0.218]	0.330 [0.251, 0.413]	0.173 [0.138, 0.210]	0.359 [0.274, 0.448]
	whole page − cited	0.099 [0.054, 0.145]	0.017 [-0.076, 0.106]	0.107 [0.063, 0.148]	-0.009 [-0.103, 0.081]
NIV	cited − uncited page	-0.004 [-0.017, 0.010]	-0.021 [-0.093, 0.059]	-0.001 [-0.013, 0.013]	-0.014 [-0.081, 0.057]
	cited − top-m ranked	-0.003 [-0.019, 0.013]	0.000 [0.000, 0.000]	-0.003 [-0.017, 0.009]	0.023 [0.000, 0.079]
	cited − random active	0.006 [-0.004, 0.020]	0.004 [-0.039, 0.069]	0.003 [-0.004, 0.012]	0.007 [-0.033, 0.067]
	whole page − cited	0.067 [0.038, 0.098]	0.050 [0.000, 0.129]	0.041 [0.016, 0.069]	0.000 [-0.068, 0.070]
Vasopressin	cited − uncited page	0.114 [0.040, 0.183]	0.163 [0.080, 0.238]	0.015 [-0.121, 0.144]	-0.063 [-0.237, 0.118]
	cited − top-m ranked	0.103 [0.034, 0.172]	-0.054 [-0.129, 0.009]	0.008 [-0.088, 0.104]	-0.204 [-0.351, -0.064]
	cited − random active	0.213 [0.149, 0.274]	0.341 [0.270, 0.404]	0.077 [-0.010, 0.165]	0.195 [0.054, 0.336]
	whole page − cited	0.305 [0.233, 0.383]	0.036 [-0.019, 0.096]	0.233 [0.142, 0.331]	0.122 [-0.028, 0.273]

Table S13: Card-intervention test of the trained TRACE explainer. The published checkpoint regenerates explanations for the frozen natural-test cohort (600 states $\times$ 2 policy targets) after the concept-card input is perturbed; the blinded Claude Opus 5 judge (matched arm, low effort) then recovers the target policy's action. *none* is a same-hardware regeneration of the unperturbed cards (the reference for the paired comparison); *random cards* replaces the page with other active but unselected dictionary concepts; *shuffled bands* keeps the concepts but permutes their activation bands; *drop top card* removes the rank-1 card.

Cohort	Cards	RL Acc	RL F1	Beh Acc	Beh F1
RRT	published	0.755	0.688	0.935	0.926
	none (regenerated)	0.752	0.687	0.935	0.926
	random cards	0.593	0.513	0.748	0.752
	shuffled bands	0.740	0.642	0.832	0.814
	drop top card	0.728	0.669	0.905	0.900
NIV	published	0.842	0.653	0.892	0.662
	none (regenerated)	0.845	0.653	0.903	0.692
	random cards	0.693	0.517	0.818	0.521
	shuffled bands	0.832	0.639	0.873	0.667
	drop top card	0.788	0.596	0.912	0.603
Vasopressin	published	0.807	0.771	0.975	0.987
	none (regenerated)	0.807	0.773	0.970	0.985
	random cards	0.732	0.673	0.955	0.977
	shuffled bands	0.802	0.767	0.970	0.985
	drop top card	0.793	0.749	0.977	0.988

Table S14: Paired change in judge accuracy / macro-F1 versus the same-hardware *none* regeneration (patient-clustered bootstrap 95% CI, 2,000 draws) and the fraction of rows whose judged action changed.

Cohort	Perturbation	ΔRL Acc	ΔRL F1	RL changed	ΔBeh Acc	ΔBeh F1	Beh changed
RRT	random cards	-0.158 [-0.205, -0.115]	-0.174 [-0.225, -0.128]	0.382	-0.187 [-0.227, -0.147]	-0.175 [-0.215, -0.138]	0.243
	shuffled bands	-0.012 [-0.051, +0.026]	-0.046 [-0.093, -0.002]	0.212	-0.103 [-0.136, -0.071]	-0.113 [-0.147, -0.079]	0.157
	drop top card	-0.023 [-0.058, +0.010]	-0.018 [-0.051, +0.013]	0.182	-0.030 [-0.047, -0.013]	-0.026 [-0.043, -0.008]	0.047
NIV	random cards	-0.152 [-0.194, -0.108]	-0.137 [-0.210, -0.066]	0.300	-0.085 [-0.119, -0.055]	-0.171 [-0.273, -0.055]	0.150
	shuffled bands	-0.013 [-0.039, +0.011]	-0.015 [-0.043, +0.009]	0.113	-0.030 [-0.055, -0.005]	-0.025 [-0.071, +0.018]	0.082
	drop top card	-0.057 [-0.089, -0.024]	-0.057 [-0.096, -0.021]	0.183	+0.008 [-0.016, +0.032]	-0.088 [-0.189, +0.024]	0.067
Vasopressin	random cards	-0.075 [-0.111, -0.042]	-0.100 [-0.147, -0.056]	0.175	-0.015 [-0.041, +0.009]	-0.008 [-0.021, +0.005]	0.062
	shuffled bands	-0.005 [-0.024, +0.014]	-0.006 [-0.028, +0.019]	0.062	+0.000 [-0.017, +0.018]	+0.000 [-0.009, +0.009]	0.043
	drop top card	-0.013 [-0.041, +0.014]	-0.024 [-0.060, +0.011]	0.097	+0.007 [-0.009, +0.023]	+0.003 [-0.005, +0.012]	0.033

Table S15: Two explanations of the same state. Both explainers see the identical patient state and never the action, and differ only in which policy's concepts accompany it. The RL policy selects IHD here and the behavior policy No-RRT; the blinded judge recovers each action from its own explanation, and CRRT for both from the state alone.

Patient state *shared by both explainers*	Age 71 years; weight 102.2 kg; heart rate 82.7 bpm; pH 7.43; blood pressure 110.3/46.3 mmHg (mean 66); SpO_2 94.7%; BUN 69 mg/dL; creatinine 4.7 mg/dL; bicarbonate 21 mEq/L; potassium 4.8 mEq/L; sodium 138 mEq/L; P/F ratio 237.5; 6-hour urine output 0 mL.
Behavior explainer *hidden action: No-RRT*	… The cohort-relative blood pressure, renal solutes, urine output, and acid-base status pattern *links* creatinine 4.7 mg/dL, mean blood pressure 66 mmHg, weight-normalized urine output 0 mL/kg/h, and bicarbonate 21 mEq/L. The … renal solutes, urine output, electrolytes, and blood pressure pattern *additionally joins* 0 mL/6h urine output with potassium 4.8 mEq/L … **These measurable joint features support serial observation and reassessment of the immediate corrective burden.**
RL explainer *hidden action: IHD*	… The cohort-relative blood pressure, renal solutes, urine output, and acid-base status pattern *emphasizes higher* creatinine *with lower* mean pressure, weight-normalized urine output, and bicarbonate; the … renal solutes, urine output, and oxygenation pattern *emphasizes higher* creatinine *with lower* urine output and P/F ratio. Creatinine is 4.7 mg/dL, urine output is 0 mL/6h … **jointly supporting the usefulness and tolerability of efficient short-interval correction.**

S1 Reporting Scope and Cohorts

This supplement documents the cohorts, clinical policies, training protocols, and evaluation procedures behind every result in the main paper, and reports additional analyses. Scores are proportions unless indicated otherwise, and a dash marks a cell that does not apply to its row.

Scope of the evidence by cohort. The three cohorts do not support every analysis equally. RRT carries the most complete evidence: explanations recover both policies' actions, stay policy specific on the disagreement states, and removing the concepts an explanation cites from the frozen representation changes the RL decision more often than removing supplied but uncited concepts (Section S3.2). On NIV the explanations are policy specific and depend on the supplied cards, but the RL policy lies far from its decision boundary, so removing concepts rarely changes its action and the policy-side test is not informative: on the 40 NIV disagreement states with a size-matched control, removing the cited concepts flips fewer RL decisions than removing size-matched uncited cards (0.025 against 0.046). On vasopressin the state-only ablation recovers the RL policy at least as well as TRACE (0.818 against 0.807; main Table 2), and the policy-side contrast depends on the citation rule. The estimated behavior policy never initiates vasopressin on the 600 vasopressin test states, so its target is a single class: answering "no vasopressin" on every state scores 1.000, and behavior-target accuracy there measures only whether an explanation avoids arguing for initiation, not explanation quality; the RL target is the informative one on that cohort.

Table S16: Cohort inventory. RRT and NIV counts are all modelled decision steps; the vasopressin count is its eligible decision states (pre-initiation epochs plus the initiation epoch). Evaluation subsets are smaller than the full cohorts.

Benchmark	Patients	States	Actions
RRT modality	2,095	114,113	No RRT; IHD; CRRT
NIV switching	2,145	18,320	Discontinue ventilation; continue NIV; intubate
Vasopressin initiation	3,102	57,574	Continue norepinephrine alone; initiate vasopressin

RRT. The source cohort contains 2,095 patients and 2,292 ICU stays. Explanations are evaluated on 600 test states from 263 held-out patients, with frozen-policy action counts 359/39/202 for No RRT/IHD/CRRT. Recovery is scored against the frozen-policy action; the observed clinician action is a separate target.

NIV analogue. The MIMIC-IV cohort is an analogue of the clinical task of Feng et al. (2023): 18,320 decision states from 2,145 patients with one ICU stay each, split patient-disjointly and stratified by 28-day mortality into 1,716, 214, and 215 training, validation, and test patients (14,726, 1,800, and 1,794 states). Observed actions are 43.7% discontinue, 54.6% continue, and 1.7% intubate. The teacher data, SFT, and all test results on this cohort use this extraction (Section S5).

Policy implementations. Each cohort uses a three-head BCQ whose shared encoder feeds a Q head, an imitation head that alone defines the support mask, and a natural-distribution clinician head trained in the same optimization step; the RL target is the masked-greedy BCQ action and the estimated behavior target is the clinician-head argmax. RRT uses a 64-dimensional encoder with a residual clinician branch. On NIV a three-layer, width-64 encoder reads the 76-input state, the imitation head is trained under the 46/46/8 resampler (support threshold 0.3), the clinician head is a residual adapter of width 16 plus a linear readout, and model selection minimizes the validation relative Bellman residual plus the weighted imitation and clinician cross-entropies (epoch 18). On the NIV test split the deployed BCQ action agrees with the recorded clinician action on 0.861 of states and with the clinician-head argmax on 0.881; the clinician head has accuracy 0.899 and one-versus-rest AUC 0.959/0.952/0.920 (discontinue/continue/intubate).

S1.1 Vasopressin Initiation: OVISS Cohort Analogue

The cohort reconstructs the MIMIC-IV cohort of the Optimal Vasopressin Initiation in Septic Shock (OVISS) study (Kalimouttou et al., 2025) (Table S17): adults meeting the official MIMIC Sepsis-3

definition at their first qualifying episode, with ongoing norepinephrine and lactate > 2 mmol/L within six hours after norepinephrine starts. Of its 3,102 patients (one septic-shock episode each), 1,535 (49.5%) have a recorded vasopressin infusion starting within the trajectory; marking initiation at the first hourly epoch with a positive-rate infusion also counts 27 patients whose infusion started before the trajectory window and continued into it (1,562 initiators, 50.4%).

Trajectory construction. Hourly epochs run from shock onset (first positive-rate norepinephrine interval overlapping the Sepsis-3 suspected-infection window) to the earliest of 12 hours after the last norepinephrine interval, ICU discharge, or 120 hours after onset (155,820 epochs); measurements are epoch-averaged, carried forward with indicators, median-imputed, and standardized on the training split. Patient splits stratified by hospital mortality give 2,481/310/311 training/validation/test patients, 46,187/6,311/5,076 decision states, 1,240/163/147 recorded initiations, and mortality 43.8%/43.9%/43.7%.

Table S17: Constructed vasopressin cohort versus the published OVISS MIMIC-IV cohort (Kalimouttou et al., 2025). Continuous values are median (IQR); initiation-time measurements are among patients with observed initiation. Percentages are recomputed from the displayed counts (denominator: the full cohort, or the initiators for RRT at initiation) and may differ from rounded percentages in the source report.

Measure	Constructed (v3.1)	Paper (v1.0)
Patients	3,102	3,056
Observed vasopressin initiation	1,535 (49.5%)	1,485 (48.6%)
Time to initiation, h	4 (1–10)	4 (1–12)
Norepinephrine at initiation, μg/kg/min	0.30 (0.18–0.42)	0.28 (0.14–0.40)
SOFA at initiation	9 (7–11)	10 (8–12)
Lactate at initiation, mmol/L	4.4 (2.8–7.2)	2.6 (1.7–5.2)
RRT at initiation	64 (4.2%)	74 (5.0%)
Hospital mortality	1,359 (43.8%)	1,186 (38.8%)

Action space and decision timing. At each decision state $\mathcal{A} = \{0, 1\}$: $a_t = 0$ continues norepinephrine alone and $a_t = 1$ initiates vasopressin in the current epoch; dose selection, titration, and weaning are outside the action space. Decision states are the epochs with active norepinephrine before vasopressin has started plus the initiation epoch (a non-initiator contributes its whole norepinephrine course with action 0, an initiator the sequence $0, \ldots, 0, 1$): 57,574 states, 1,550 (2.7%) of them recorded initiations. Post-initiation epochs and epochs without active norepinephrine remain in the trajectories as ordinary transitions, so later rewards and the terminal outcome are credited to earlier decisions, but every explanation and evaluation result on this cohort uses decision states only.

Policy state. The 43-column state holds 17 hourly variables (norepinephrine dose, SOFA, lactate, mean arterial pressure, heart rate, systolic blood pressure, respiratory rate, temperature, creatinine, blood urea nitrogen, hemoglobin, platelets, white-cell count, INR, total bilirubin, Glasgow Coma Scale, weight-normalized urine output), each with its observation indicator; mechanical-ventilation and kidney-replacement-therapy indicators; age, weight, Charlson index, and sex; hours since shock onset; and hours since initiation and whether vasopressin was running at epoch start (lagged one epoch, so the state never contains the current action). Unlike the reference inventory (Kalimouttou et al., 2025), it omits hourly glucocorticoid exposure and intravenous-fluid volumes.

Policy. We train a three-head BCQ (three-layer, width-64 encoder; support threshold 0.3; width-16 residual clinician adapter fitted on natural-distribution decision states). As initiation is only 2.7% of decision states, batches draw decision states with equal action shares (other epochs keep their natural share) and the temporal-difference term carries importance weights $(p/q)^{0.4}$ (natural over sampled share), normalized to a maximum of one. Training uses the reward of Equation 3, $\gamma = 0.99$, and a hard target update every 10 steps; the selected epoch (39) minimizes normalized validation temporal-difference error plus imitation and clinician losses. On the 5,076 test decision states the deployed action initiates on 33.4% and agrees with recorded actions on 68.2%. The clinician head ranks initiations well (AUROC 0.880) and is calibrated in the mean (0.030 predicted, 0.029 recorded), but

its initiation probability exceeds 0.5 on only 2 of these states, so the estimated behavior target almost never initiates (Table S7).

Definition sensitivity and replication scope. The OVISS report does not specify a lactate-to-vasopressor timing window; the six-hour requirement is our reconstruction choice (wider windows produced 4,326–4,561 patients). Differences from the published cohort in shock-onset SOFA, Glasgow Coma Scale, lactate, and mechanical-ventilation prevalence, together with the release differences, suggest timing or coding differences that the published description does not resolve, so similar cohort size is not evidence of exact replication; we treat the cohort as a version-aware benchmark analogue with a documented extraction specification.

S1.2 Clinical Policy Rewards

The clinical rewards r_t^{clin}, distinct from the explanation-generation reward used by PPO, follow the studies cited for each cohort and train the RL branch; the behavior branch is estimated from recorded clinician actions by cross-entropy, and both clinical policies are frozen before explainer training.

RRT: survival and SOFA shaping. Following the mixed reward in Equations (2)–(3) and the calibrated setting in Table 2 of Xu et al. (2026), the RRT policy uses

$$r_t^{\text{RRT}} = \begin{cases} +a, & d_t = 1,\ m_{\text{ICU}} = 0, \\ -a, & d_t = 1,\ m_{\text{ICU}} = 1, \\ -c(S_{t+1} - S_t) - b\,\mathbb{I}\{S_{t+1} = S_t > 0\}, & d_t = 0, \end{cases} \tag{1}$$

where d_t marks the final transition of a trajectory, m_{ICU} indicates ICU death, S_t is the Sequential Organ Failure Assessment (SOFA) score, $a = 10$, $b = 0.033$, $c = 0.25$, and $\gamma = 0.99$: a SOFA decrease earns a positive intermediate reward, an unchanged positive SOFA a small penalty, and the terminal reward replaces the intermediate term on the final transition.

NIV: 28-day outcome with potential-based shaping. Following the 28-day mortality objective of Feng et al. (2023), the NIV reward combines a sparse terminal outcome with two dense terms:

$$r_t^{\text{NIV}} = r_t^{\text{mort}} + \lambda_\Phi \big[\gamma\,\Phi(s_{t+1}) - \Phi(s_t)\big] + \lambda_{\text{free}}\, R^{\text{free}}(s_{t+1}), \quad r_t^{\text{mort}} = \begin{cases} +10, & d_t = 1,\ m_{28} = 0, \\ -10, & d_t = 1,\ m_{28} = 1, \\ 0, & d_t = 0, \end{cases} \tag{2}$$

where m_{28} indicates death within 28 days. The respiratory-recovery potential $\Phi(s) = -\text{logit}\,\widehat{P}(\text{intubation or death within 24 h of the bin end} \mid s)$ comes from an L_2 logistic model on the leakage-free physiological state (25 features and elapsed time, no availability masks), fitted on the training split (cross-fitted across five subject-level folds for training rows) and standardised with training statistics; $\Phi = 0$ at the absorbing terminal, so the shaping is potential-based and leaves the optimal policy unchanged. $R^{\text{free}}(s_{t+1})$ is $+1$ when the patient is alive and off noninvasive and invasive ventilation after the transition, 0 on noninvasive ventilation, and -1 on invasive ventilation; $\lambda_\Phi = 0.5$, $\lambda_{\text{free}} = 0.1$, and $\gamma = 0.99$.

Vasopressin: OVISS composite reward. We adopt the reward specification in Supplement 1 (eMethods, Reward function) of Kalimouttou et al. (2025):

$$r_t^{\text{OVISS}} = 1 - 20 d_t m_{\text{hosp}} + I_t^{\text{Lac},6} + I_t^{\text{MAP},4} + 3 I_t^{\text{SOFA},6} + I_t^{\text{NE},4}, \tag{3}$$

where d_t marks the final epoch of a trajectory and m_{hosp} indicates hospital death, so $r_t \in [-19, 7]$. Each indicator looks only at later epochs and fires only on values measured there, before any imputation: $I_t^{\text{Lac},6} = 1$ when a lactate measured in epochs $t+1, \ldots, t+6$ is lower than the current value; $I_t^{\text{MAP},4} = 1$ when the current mean arterial pressure is below 65 mmHg and a value measured in $t+1, \ldots, t+4$ is at least 65 mmHg; $I_t^{\text{SOFA},6} = 1$ when a SOFA score recomputed from new measurements in $t+1, \ldots, t+6$ is lower than the current score; and $I_t^{\text{NE},4} = 1$ when the norepinephrine rate in effect in any of $t+1, \ldots, t+4$ is lower than the current rate, counting a stopped infusion as zero. Near the end of a trajectory a window covers only the remaining epochs (8.0% and

11.9% of epochs have a truncated four- and six-hour window), and a window without a qualifying measurement scores 0. The terminal transition is not bootstrapped, and future measurements and mortality enter only the reward, never a decision-time input.

S2 TEACHER DATA, SFT, AND SHARED PPO PROTOCOL

TRACE trains one explainer per cohort, shared by its RL and estimated behavior targets. Each training record carries a policy indicator p, visible to the student and reader, selected cards $C_p(s)$, and a private target action $a^p(s)$ used only for teacher conditioning, filtering, and reward calculation. All records from a patient stay in one split, whichever policy they explain.

Teacher and SFT. On RRT the card-aware teacher, GPT-5.6 Terra, receives training-state measurements, eight selected cards, and a private action-conditioned objective; its output contract requests 55–135 words, at most four sentences, two to four named cards, and at least two matching patient measurements. The student sees the state, policy indicator, and cards, not the objective. The paired teacher data hold 19,096 records (9,548 state–policy pairs); 17,850 enter each training view once development patients are held out. The evaluated SFT checkpoint continues a three-epoch RL-target SFT run for two epochs on this pool, with target-policy and action-balanced sampling in training only. NIV and vasopressin follow the same contract with their own teacher data (Sections S5 and S6).

Explanation reward, shared by the three cohorts. All cohorts optimize main-paper Equation 3 with the same implementation and weights: recovery $\lambda_r = 1.0$, concept (card attribution) $\lambda_c = 0.3$, grounding $\lambda_g = 0.2$, leakage $\lambda_\ell = 1.0$, and a KL penalty ($\beta = 0.05$) subtracted per response token before advantage estimation. The recovery term splits an explanation into at most six cumulative sentence prefixes, the last being the whole explanation, and places the increment in the frozen Qwen3-8B reader's target-action log probability at each prefix's final response token, the first increment measured from a state-only query (probabilities floored at 10^{-8}). The sentence-prefix increments telescope: their sum equals the reader's log-probability gain for the target action from the state-only query to the complete explanation. The decomposition therefore does not credit sentences independently; it distributes that gain over the response at sentence boundaries, which gives PPO a denser signal than a single terminal reward. The concept term is $(\mathbb{I}\{\text{frame}\} + \mathbb{I}\{\text{evidence}\}) - 1 \in \{-1, 0, 1\}$: the explanation must open with the selected-representation phrasing in cohort-relative language (frame) and cite at least two patient values listed on the named selected cards (evidence); a record without selected cards scores -1. The grounding term is $+1$ when every stated number is supported by the reference text (display-rounding tolerance $\max(0.051,\ 0.005\,|v|)$ per value v), -1 when any is not, and 0 when none is stated. The concept and grounding terms are deterministic checks of form and numeric consistency (the opening framing, at least two card-listed values, every number matching the reference text). They do not test whether the explanation states a concept's clinical meaning or direction correctly. The leakage indicator fires when an explanation names an action or intervention, uses a reader label letter, or identifies which policy it explains; these three terms sit at the final response token. The w/o concepts arm (no cards) sets $\lambda_c = 0$ and keeps every other weight. Only the cohort interface differs: the reader's action set, label permutations, cards per record, grounding reference (Table S18), and action-leakage lexicon, the last copied from the cohort's teacher acceptance rule. NIV and vasopressin, as in that rule, also accept card-displayed values; NIV also admits the GCS total derived from its three component scores, and vasopressin credits a card's $0/1$ axis (measurement availability, mechanical ventilation, kidney replacement therapy, prior vasopressin exposure) when the explanation names it, since a quarter of its card pages lack a numeric axis.

Shared PPO. One actor LoRA and one value head are optimized over the pooled targets; the two clinical policies, SAE, dictionary, SFT backbone, reader, and SFT reference are fixed. Since both targets share reader and reward, an RL–behavior difference reflects evidence and target action, not supervision; recovery is monitored per policy so an aggregate gain cannot hide degradation on one target. All three cohorts use LoRA rank 16 and scaling 32, PPO clip 0.2, value-loss coefficient 0.1, discount 1.0, GAE parameter 0.95, linear learning-rate decay, rollout batch 12, minibatch 4, four inner epochs, sampling temperature 0.7, at most 384 new tokens, and seed 200. RRT and vasopressin use two passes over the paired pool at learning rate 2×10^{-5}; on NIV, a second round

Table S18: Cohort-specific parts of the shared PPO explanation reward; the reward weights and β are identical across cohorts.

	RRT	NIV	Vasopressin
Actions scored by the reader	3	3	2
Label permutations averaged	6	6	2
Concept cards per record	8	8	4
Grounding reference	state	state + cards	state + cards
Prompt cap (tokens)	3,072	6,144	5,632

continued from the first round's final checkpoint for two more passes at 1×10^{-5} (2,780 iterations in total; checkpoint of record chosen over both rounds, Section S5). Each pair-aligned rollout batch holds six states with both targets, so every minibatch carries both and the loss $\frac{1}{2}L_{\mathrm{RL}} + \frac{1}{2}L_{\mathrm{Behavior}}$ fixes $\rho_{\mathrm{RL}} = \rho_b = 1/2$ in Equation 4. States are balanced over the learned policy's actions; within each action class a state comes from the class's disagreement states with probability $\min(0.5,\, 3f_a)$, f_a being their share of the class, and otherwise from the whole class; sampling is not adaptive.

Final judge, accuracy, and checkpoints of record. Claude Opus 5, separate from the reader that supplies PPO rewards, predicts the target policy's action from an explanation, the decision context, and the candidate action definitions; the correct action, concept cards, Q values, and policy probabilities are withheld. Accuracy, the fraction of all evaluated cases whose prediction equals the frozen policy's action, is computed separately for the RL and estimated behavior targets. A malformed or out-of-set reply is re-requested up to three times (the run aborts if all fail), so every row carries a valid label; schema-invalid or truncated explanations stay in the denominator. Candidate definitions, context visibility, label mapping, decoding, and parser are fixed across methods; macro-F1 and per-action recall accompany accuracy, and patient-cluster resampling quantifies uncertainty. Recovery credits the whole explanation, including how it phrases what its cited evidence implies for the decision: with that phrasing removed and every named concept and cited measurement kept, recovery falls substantially on RRT and NIV and slightly on vasopressin, and where PPO improves on Concept SFT, most of the improvement lies in that phrasing. Whether the cited concepts carry the decision is tested separately, by intervention on the frozen policies (Section S3.2). All main-table methods are scored this way on the frozen test cohorts.

S3 Additional Evaluations on the Held-Out Test Cohorts

This section reports the RRT checkpoint comparison with its matched-evaluation controls, concept interventions on the frozen policies, and the paired RL–behavior concept comparison, all on the frozen 600-state test cohorts.

Table S19: Action-recovery accuracy of every cell of main Tables 1 and 2 with its patient-clustered bootstrap 95% interval (2,000 resamples over the 263, 194, and 166 test patients of the RRT, NIV, and vasopressin cohorts; 600 states per cell). Rows above the rule are main Table 1, rows below it main Table 2. †Judged against the recorded clinician action on a 294-state action-balanced subset. The vasopressin behavior target is a single class on the other rows (Section S1).

	RRT		NIV		Vasopressin	
Method	RL	Behavior	RL	Behavior	RL	Behavior
Direct LLM	0.437 [0.391, 0.482]	0.492 [0.442, 0.540]	0.500 [0.443, 0.557]	0.622 [0.564, 0.682]	0.678 [0.633, 0.720]	0.675 [0.621, 0.726]
Zero-shot CoT	0.438 [0.392, 0.484]	0.457 [0.402, 0.510]	0.512 [0.455, 0.568]	0.638 [0.581, 0.696]	0.720 [0.674, 0.762]	0.862 [0.820, 0.898]
MCTS Refinement	0.373 [0.330, 0.418]	0.428 [0.377, 0.481]	0.528 [0.478, 0.580]	0.647 [0.591, 0.703]	0.500 [0.447, 0.556]	0.807 [0.759, 0.852]
BC-LLM	0.487 [0.442, 0.532]	0.608 [0.562, 0.652]	0.502 [0.457, 0.546]	0.592 [0.546, 0.638]	0.617 [0.574, 0.660]	0.684† [0.631, 0.737]
Policy distillation + LLM	0.412 [0.366, 0.456]	0.423 [0.369, 0.476]	0.537 [0.483, 0.590]	0.633 [0.578, 0.691]	0.683 [0.627, 0.743]	0.888 [0.844, 0.927]
SFT	0.697 [0.659, 0.732]	0.873 [0.843, 0.903]	0.545 [0.498, 0.594]	0.578 [0.533, 0.626]	0.765 [0.729, 0.801]	0.957 [0.936, 0.975]
CoT + SFT	0.670 [0.630, 0.709]	0.855 [0.823, 0.884]	0.535 [0.492, 0.583]	0.593 [0.550, 0.637]	0.772 [0.734, 0.806]	0.960 [0.937, 0.978]
PPO + BT reward	0.478 [0.431, 0.527]	0.483 [0.433, 0.532]	0.618 [0.571, 0.664]	0.705 [0.655, 0.754]	0.700 [0.650, 0.746]	0.933 [0.904, 0.958]
RLAIF (CNF rewards)	0.475 [0.424, 0.522]	0.490 [0.437, 0.544]	0.517 [0.461, 0.574]	0.625 [0.567, 0.686]	0.773 [0.734, 0.812]	0.900 [0.864, 0.931]
TRACE (Ours)	0.755 [0.716, 0.792]	0.935 [0.909, 0.958]	0.842 [0.809, 0.873]	0.892 [0.862, 0.919]	0.807 [0.772, 0.841]	0.975 [0.956, 0.990]
w/o PPO	0.742 [0.703, 0.776]	0.897 [0.869, 0.923]	0.705 [0.665, 0.744]	0.788 [0.747, 0.829]	0.800 [0.762, 0.836]	0.917 [0.885, 0.945]
w/o concepts	0.612 [0.571, 0.650]	0.838 [0.806, 0.870]	0.532 [0.490, 0.576]	0.637 [0.591, 0.685]	0.818 [0.786, 0.849]	0.983 [0.968, 0.995]
w/o behavior policy	0.773 [0.735, 0.808]	–	0.852 [0.822, 0.881]	–	0.820 [0.782, 0.854]	–

Table S20: What the judge recovers with and without TRACE's explanation, on the 600-state test cohort of each benchmark. *Matched*: the state, the policy tag, and the policy's own explanation; *state only*: the same state and tag without an explanation; *shuffled*: an explanation written for a different state with a different hidden action (not constructible for the one-class vasopressin behavior target); *other policy*: the other policy's explanation of the same state under the same tag. Brackets: patient-clustered bootstrap 95% intervals of the paired differences.

		All 600 states				Disagreement states				
Cohort	Target	Matched	State only	Shuffled	Matched − state only	n	Matched	State only	Other policy	Matched − other
RRT	RL	0.755	0.453	0.113	+0.302 [0.249, 0.354]	119	0.706	0.445	0.101	+0.605 [0.504, 0.702]
	Behavior	0.935	0.488	0.085	+0.447 [0.383, 0.510]	119	0.899	0.387	0.193	+0.706 [0.615, 0.793]
NIV	RL	0.842	0.548	0.125	+0.293 [0.238, 0.349]	76	0.513	0.211	0.000	+0.513 [0.386, 0.632]
	Behavior	0.892	0.625	0.190	+0.267 [0.219, 0.314]	76	1.000	1.000	0.474	+0.526 [0.403, 0.644]
Vasopressin	RL	0.807	0.690	0.307	+0.117 [0.067, 0.171]	206	0.602	0.534	0.150	+0.451 [0.355, 0.551]
	Behavior	0.975	0.958	–	+0.017 [−0.008, 0.040]	206	0.961	0.903	0.432	+0.529 [0.449, 0.613]

What the judge can recover without the explanation. In every arm the judge sees the patient state and the policy tag, so part of the target action may be recoverable without any explanation. The state-only control supplies both and omits the explanation, so matched minus state-only accuracy is the recovery the explanation adds (Table S20). On the RL target, TRACE's explanations add $+0.302$ $[+0.249, +0.354]$ on RRT and $+0.293$ $[+0.238, +0.349]$ on NIV over the state-only control; on vasopressin the addition is smaller (matched 0.807 against state-only 0.690, $+0.117$ $[+0.067, +0.171]$) and is not resolved on its 206 disagreement states ($+0.068$ $[-0.034, +0.171]$; matched 0.602 against state-only 0.534). Replacing the explanation with one written for a different state with a different hidden action (the shuffled control) drives recovery far below the state-only level (0.113 on the RRT RL target), so the judge follows the explanation it is given rather than the state. On the RRT disagreement states, the other policy's explanation of the same state likewise leads the judge to the other policy's action (Table S9). Table S19 gives every accuracy cell of main Tables 1 and 2 with its patient-cluster 95% interval; these intervals cover patient resampling only, not training-seed or judge run-to-run variation.

S3.1 Shared PPO on the Paired RRT Test Cohort

The run contains 2,975 PPO iterations; the checkpoint of record is iteration 2604, selected on development recovery accuracy. Table S21 compares it with its Concept SFT parent on the same 600 patient states from 263 held-out patients, with one explanation per target policy (1,200 rows per judge arm), deterministic generation, and a 384-new-token limit. This comparison describes one training run.

Controls and uncertainty. The frozen Claude Opus 5 judge (low effort) sees the patient state, target-policy indicator, candidate action definitions, and the assigned explanation. The four arms are matched explanation, cross-policy explanation for the same state (the target-policy indicator stays fixed), state only, and a targeted shuffled explanation drawn with replacement from explanations with the same policy tag but a different hidden action; the shuffle is an action-mismatched control, not an unrestricted random shuffle. Candidate-label mappings follow the same deterministic protocol for both checkpoints, and each arm retains all 1,200 rows. Intervals use 2,000 patient-cluster bootstrap resamples; the intervals for matched-minus-cross quantify a within-checkpoint contrast, not the significance of the PPO-minus-SFT improvement.

Within-state paired policy recovery. Let D be the 119 states for which $a^{\mathrm{RL}}(s) \neq a^{b}(s)$, and let $J(s, p, e)$ be the action decoded from the judge when supplied state s, policy tag p, and explanation e. We measure whether both matched queries recover their respective policy actions:

$$\frac{1}{|D|} \sum_{s \in D} \mathbb{1}\big[J(s, \mathrm{RL}, e_{\mathrm{RL}}(s)) = a^{\mathrm{RL}}(s)\big]\, \mathbb{1}\big[J(s, b, e_b(s)) = a^{b}(s)\big]\,. \tag{4}$$

Both policy rows of a state are verified to belong to the same patient and to have different reference actions, and failures are retained as incorrect. Table S8 partitions the states into four mutually exclusive outcomes. At iteration 2604, both actions are recovered on 74/119 states (0.622; patient-cluster 95% interval [0.528, 0.711]), versus 53/119 under Concept SFT (0.445; [0.357, 0.535]). The

Table S21: Concept SFT versus the checkpoint of record on the paired RRT test cohort under Opus 5. Matched, cross, shuffled, and macro-F1 use all 600 states per policy. The final column is matched-minus-cross accuracy on the 119 disagreement states, with 95% patient-cluster bootstrap intervals. RL macro-F1 uses three supported classes; behavior uses two (No RRT and CRRT).

Checkpoint	Policy	Matched	Cross	Shuffled	Macro-F1	Gap on disagreement
Concept SFT	RL	0.742	0.765	0.147	0.672	+0.378 [0.261, 0.496]
TRACE (2604)	RL	0.755	0.777	0.113	0.688	+0.605 [0.504, 0.702]
Concept SFT	Behavior	0.897	0.698	0.107	0.908	+0.437 [0.322, 0.545]
TRACE (2604)	Behavior	0.935	0.652	0.085	0.926	+0.706 [0.615, 0.793]

paired PPO-minus-SFT difference is 0.176 [0.094, 0.262]. These intervals (2,000 patient-cluster resamples, seed 42) preserve the two policy queries and the checkpoint pairing within each state and capture evaluation-sample uncertainty only.

The swap matrix of Figure S2 instead reports separate queries: each column fixes both the judge's policy tag and the reference action while the explanation source changes, so its rows need not sum to one and it is not a confusion matrix from a policy-blind judge. The paired measure above changes both the explanation source and the judge tag between its two queries and should not be interpreted as an explanation-only intervention.

Policy specificity and class imbalance. From Concept SFT to iteration 2604, the matched-minus-cross gap on disagreement states increases from 0.378 to 0.605 for RL and from 0.437 to 0.706 for behavior. Full-set RL cross accuracy exceeding matched accuracy is driven by agreement states: cross recovery is 0.944 there but only 0.101 on disagreement states, where it follows the other policy's action in 0.899 of cases (Table S9), so full-set cross accuracy alone does not measure policy specificity. Matched macro-F1 (0.688 for RL, 0.926 for behavior) averages different class sets, because the behavior target contains no IHD decisions, and the two values must not be compared directly. The main remaining error is IHD over-prediction: the RL judge predicts IHD 124 times for 39 true IHD targets (recall 0.949, precision 0.298), whereas CRRT precision and F1 reach 0.879 and 0.814, versus 0.817 and 0.794 under Concept SFT, with recall 0.757 versus 0.772.

S3.2 Concept Interventions on the Frozen Policies

Removing concepts from the frozen representation and re-running the complete decision rule gives cohort-specific evidence. On RRT, removing the concept page flips the deployed RL action on 41% of test states, against 20% for size-matched random active concepts and 5% for the smallest-effect concepts, and removing only the two or three concepts an explanation cites flips it on 61.3% of disagreement states, against 40.2% for size-matched uncited page cards. On NIV the policy sits far from its decision boundary, and the test is uninformative. On vasopressin only six coordinates are active per state, and the page-level test does not discriminate. On the explainer's side, replacing the policy-selected cards with a size-matched set of unselected concepts (active for the state, except two of the four replacements on vasopressin) lowers RL recovery by 0.158, 0.152, and 0.075 on RRT, NIV, and vasopressin.

Protocol. For each state of the frozen 600-state test cohort and a concept set S, we ablate the concepts from the true representation, $h_S = h - \sum_{k \in S} z_k d_k$ (decoder columns mapped back to h space on vasopressin, whose SAE is fitted in a whitened metric), and re-run the complete frozen decision rule, including the BCQ support mask from the imitation head. This is the operation the concept-selection stage uses to rank single concepts. The conditions are the *concept page* shown to the explainer ($k = 8$, 8, and 4 cards on RRT, NIV, and vasopressin), the top-k concepts by unsigned decision effect, the top-k *supportive* concepts of the chosen action (signed effect), k *random* active concepts (20 draws per state), and the k concepts with the smallest effect. Outcomes are the flip rate of the deployed action, the Q*-preference reversal* rate (the Q head re-decided over the baseline support set, on the Q-decided states whose baseline support holds at least two actions: 291 of 600 on RRT and 127 on NIV), and the fraction of the baseline Q margin removed. Intervals are patient-clustered bootstraps (2,000 draws).

Selection, circularity, and off-manifold edits. The concept page is chosen by its effect on the decision, so removing it is expected to change the decision more than removing random concepts; the page-level contrast is a consistency check of the selector, not independent evidence that the explanation is faithful. The tests that hold the selector fixed are the cited-concept ones below: removing the concepts an explanation's text cites is compared with removing equally many uncited cards of the same page and equally many of the page's top-ranked cards. On the RRT disagreement states, citation adds $+0.227$ $[+0.130, +0.321]$ over the selector's own top-ranked cards and $+0.212$ $[+0.109, +0.319]$ over uncited cards, so what the text cites carries information the ranking does not; on all RRT states it adds nothing over the ranking ($+0.005$), and on NIV and vasopressin the test is not informative. Each removal subtracts the concepts' decoder contributions from the true representation rather than decoding through the SAE, which sets those coordinates of the sparse code to zero and keeps the SAE's reconstruction residual, so reconstruction error is never attributed to the removed concepts. Every control removes the same number of active coordinates from the same representation, so an effect of leaving the data distribution that is common to all removals raises every absolute flip rate without producing the cited-minus-control differences. The sets are not matched on the size of the edit, $\| \sum_{k \in S} z_k d_k \|$, however, so an off-distribution effect that grows with edit size could add to those differences to the extent that the cited concepts are more strongly active than their controls; the contrast against the top-ranked cards, which the ranking favors partly for their contribution to the representation, is the least exposed to it.

Table S22: Concept intervention on the frozen RL policy (RRT and NIV; vasopressin in the text). On each of the 600 test states, k concepts are removed from h and the complete BCQ decision rule (Q head and imitation-head support mask) is re-run. *Flip*: the deployed action changes. *Q-preference reversal*: the edited Q head, over the *baseline* support set, changes its choice (states whose baseline support holds at least two actions). *Margin removed*: change in the Q margin between the chosen action and its runner-up as a fraction of the baseline margin, on the same states. *Concept page (k cards)* is the full page shown to the explainer ($k = 8$ on both cohorts), not the subset an explanation's text cites, which is analysed below; the other four rows are controls. Random removal averages 20 draws per state; brackets are patient-clustered bootstrap 95% intervals.

Benchmark	Removed concepts	Flip	Q-preference reversal	Margin removed
RRT	Concept page (k cards)	0.405 [0.362, 0.451]	0.495 [0.433, 0.556]	+1.26 [+0.90, +1.63]
	Top-k |effect|	0.400 [0.356, 0.445]	0.474 [0.410, 0.540]	+1.19 [+0.85, +1.54]
	Top-k supportive	0.822 [0.789, 0.855]	1.000 [1.000, 1.000]	+4.11 [+3.92, +4.27]
	Random k active	0.201 [0.183, 0.219]	0.245 [0.220, 0.269]	+0.11 [+0.07, +0.16]
	Bottom-k effect	0.050 [0.032, 0.068]	0.048 [0.021, 0.078]	-0.00 [-0.00, +0.00]
NIV	Concept page (k cards)	0.077 [0.054, 0.102]	0.000 [0.000, 0.000]	-0.07 [-0.13, -0.02]
	Top-k |effect|	0.077 [0.054, 0.103]	0.000 [0.000, 0.000]	-0.07 [-0.13, -0.02]
	Top-k supportive	0.103 [0.075, 0.135]	0.205 [0.128, 0.293]	+0.69 [+0.64, +0.75]
	Random k active	0.027 [0.018, 0.038]	0.001 [0.000, 0.002]	-0.01 [-0.02, +0.00]
	Bottom-k effect	0.007 [0.002, 0.014]	0.000 [0.000, 0.000]	-0.00 [-0.00, +0.00]

Flip versus Q-preference reversal. *Flip* lets both stages of the decision move, the imitation-head support mask and the Q argmax within it, whereas *Q-preference reversal* holds the support set at its baseline and asks only whether the Q ordering among the originally admissible actions reverses. *Margin removed* can move without the ordering reversing.

What the interventions show. On RRT, removing the concept page flips the deployed action on 0.405 of states, against 0.201 for random removal and 0.050 for the smallest-effect concepts, and reverses the Q ordering on 0.495 of Q-decided states against 0.245 (Table S22). Removing the eight most supportive concepts flips 0.822 of states, and every one of the 119 disagreement states, with 96% of those flips landing on the behavior policy's action; on the RRT behavior branch, removing its own page flips the clinician-model action on 30% of states against 10% for random removal. On NIV the policy sits far from its decision boundary: the mean Q-decided margin is 0.76, and only 127 of the 600 states are decided by the Q head (the support mask admits a single action on the rest). Page removal still roughly triples the random flip rate (0.077 against 0.027; paired difference $+0.050$ $[+0.031, +0.069]$), but on the 76 disagreement states page minus random

is $+0.003$ $[-0.019, +0.035]$; the supportive removal still separates there, reaching 24% against 4% for random. On vasopressin, with six active coordinates per state, removing four of them changes the deployed action often under most rules, including the four smallest-effect concepts (flip 0.520 against 0.807 for the page and 0.547 for random), and the page does not reverse the Q ordering more often than random removal (0.161 against 0.357), so the page-level test is not informative on this cohort; the signed supportive set reverses the Q ordering on 0.976 of Q-decided states against 0.357 for random removal. On the vasopressin behavior branch, removing its own page flips the clinician-model action less often than random removal (-0.099 $[-0.112, -0.087]$; -0.160 $[-0.181, -0.139]$ on disagreement states), so the page-level test gives no support there.

Two qualifications apply. The page is selected on unsigned effect, so it is the set the decision is most *sensitive* to in either direction rather than a one-sided support set; the signed supportive set is the one for a necessity claim. And a flip shows that the removed concepts carry the frozen policy's computation, not that the judge uses their natural-language rendering, which the explanation-level controls (cross-policy, shuffled, and state-only arms) address.

Intervening on the evidence the explanation asserts. To test the evidence an explanation asserts rather than the whole page, we recover, for every deployed explanation, which supplied cards its text claims and repeat the removal on that set alone. The matcher is model-free and is scored against the teacher's recorded card use in the teacher corpora (Table S23): on RRT it reaches precision 0.936 and recall 0.623 under the strict rule and F1 0.778 under the coverage rule, but strict recall is only 0.157 on NIV. The teacher's output contract asks for two to four named cards, but under the strict rule the matcher resolves on average 2.5 cited cards per eligible RRT state, 1.2 on NIV, and 1.0 on vasopressin, and the four-card vasopressin page often leaves no size-matched uncited complement. Because of matcher error and contamination of the uncited pool with cards the explanation did use, the contrasts below are likely attenuated toward zero.

On RRT the cited concepts carry the RL decision (strict rule in Table S11; paired contrasts under both rules in Table S12). Removing only the cited concepts flips the deployed RL action on 0.301 of states, against 0.215 for an equally large set of supplied but uncited cards ($+0.087$ $[+0.042, +0.128]$); on the disagreement states the cited set flips 0.613 against 0.402 ($+0.212$ $[+0.109, +0.319]$) and is statistically indistinguishable from removing the entire eight-card page (0.630; $+0.017$ $[-0.076, +0.106]$). Against the equally large set of top-ranked page cards, citation adds nothing on all states ($+0.005$ $[-0.041, +0.045]$) but $+0.227$ on the disagreement states; the coverage rule reproduces the cited-minus-uncited contrasts ($+0.093$ and $+0.290$). On the RRT behavior branch the contrast is resolved on the decision margin ($+0.158$ $[+0.080, +0.238]$) but not on the flip rate ($+0.035$ $[-0.003, +0.072]$). The other two cohorts do not support the test. On the 40 eligible NIV disagreement states the cited set flips 0.025 against 0.046 for uncited cards, and removing the whole page flips only 0.083 of the eligible RL decisions, so NIV cannot answer the question. On vasopressin the contrast is positive under the strict rule on all states ($+0.114$ $[+0.040, +0.183]$) but reverses on disagreement states under the coverage rule (-0.063 $[-0.237, +0.118]$; against top-ranked cards -0.204 $[-0.351, -0.064]$), so we make no cited-concept claim there.

Do the behavior policy's concepts carry the behavior channel of the BCQ rule? The frozen RL rule has a value channel (the action-value head) and a behavior channel (the support mask built from the imitation head); on RRT, the only cohort on which this test was run, we ask which channel each policy's page carries. We remove from h the concepts on the behavior policy's page that are not on the learned policy's page (5.3 per state on average), re-run the complete learned-policy rule, and compare with the equally large set of the learned policy's own exclusive concepts, each as a paired difference against size-matched random active concepts (20 draws per state, patient-clustered 95% intervals). The behavior-exclusive concepts change the support set ($+0.154$ $[+0.119, +0.189]$) and shift the imitation logits (maximum absolute change, $+0.242$ $[+0.215, +0.270]$), but leave the value channel untouched: with the support held at baseline, the Q-preference reversal on the 291 Q-decided states differs from random by -0.008 $[-0.049, +0.033]$. The learned-policy exclusive concepts do the opposite, reversing the Q preference ($+0.187$ $[+0.118, +0.257]$) while changing the support set only $+0.048$ $[+0.017, +0.080]$ more than random. Because the behavior page was selected on the clinician branch and never on the imitation head, the transfer to the support mask is a property of the shared representation rather than of the selection rule. The dissociation holds only for the concepts the two pages do not share: removing the full eight-card behavior page reverses

Table S23: How well the citation matcher recovers the cards an explanation asserts, and what it finds on the deployed generations. The model-free matcher reads each card's (axis, direction) claims and its name, finds axis mentions in the text by longest alias, and assigns them to cards. COVERAGE: a named card is cited when the text mentions every axis it asserts and contradicts none of its directions; STRICT: the card's name is written out, or the card wins a contiguous run of directional claims in a greedy assignment. Both rules are scored against the card ranks the teacher recorded as used for every accepted explanation in the teacher corpora (19,096 / 8,916 / 8,000 records). *Contamination*: share of cards in the predicted-uncited pool that the teacher had declared (the impurity of the control arm); it and the false positives push the cited-minus-uncited contrast toward zero. Right-hand block, deployed TRACE generations: mean cited-set size $\bar{m}$ out of k, and the share of the 1,200 explanations for which no card could be resolved.

		vs teacher's recorded card use				deployed TRACE generations		
Cohort	Rule	Prec.	Rec.	F_1	Contam.	$\bar{m}$ (RL)	$\bar{m}$ (Beh.)	no citation
RRT ($k = 8$)	STRICT	0.936	0.623	0.748	0.163	2.52	2.66	0.003
	COVERAGE	0.752	0.806	0.778	0.102	2.89	3.56	0.010
NIV ($k = 8$)	STRICT	0.889	0.157	0.267	0.376	0.63	0.51	0.520
	COVERAGE	0.775	0.290	0.422	0.347	0.90	1.61	0.355
Vasopressin ($k = 4$)	STRICT	0.959	0.180	0.303	0.776	0.44	0.36	0.610
	COVERAGE	0.823	0.763	0.792	0.745	2.69	3.24	0.084

the learned policy's Q preference as often as removing its own page ($+0.267$ $[+0.202, +0.333]$ against $+0.241$ $[+0.176, +0.308]$ over random), while still changing the support set more ($+0.310$ $[+0.270, +0.350]$ against $+0.129$ $[+0.092, +0.164]$).

Interventions on the explainer's input. To test whether the trained explainer's recovery depends on the policy-selected cards, the published TRACE checkpoint regenerates the frozen 600-state test cohort after the card input is perturbed, and the same blinded Opus 5 protocol re-judges it: *random cards* replaces the page with a size-matched set of other concepts that are active for the state but were not selected (on vasopressin, whose native code has only six active coordinates, two of the four replacements are inactive bank concepts rendered at zero activation); *shuffled bands* keeps the same concepts and permutes their activation bands and directions among them; *drop top card* removes the rank-1 card. The renderer was verified to reproduce every stored prompt byte for byte before any generation, and because greedy generation is not bit-reproducible across GPU types, the paired reference (*none*) is a same-hardware regeneration of the unperturbed cards, which reproduces the published cells within noise (Table S13; paired changes with patient-clustered intervals in Table S14).

Replacing the policy-selected cards with other active concepts is the perturbation that hurts: RL-target accuracy falls by 0.158 $[0.115, 0.205]$ on RRT, by 0.152 $[0.108, 0.194]$ on NIV, and by 0.075 $[0.042, 0.111]$ on vasopressin, and the behavior target falls by 0.187 and 0.085 on RRT and NIV, while the single-class vasopressin behavior target barely moves (-0.015 $[-0.041, +0.009]$). Scrambling the bands among the same concepts costs little on the RL target (only the RRT macro-F1 change excludes zero) but 0.103 $[0.071, 0.136]$ on the RRT behavior target; dropping the top card is largely absorbed by the remaining page on RRT and vasopressin and costs 0.057 $[0.024, 0.089]$ on the NIV RL target. These tests show that recovery depends on the supplied cards; they do not show that every cited concept, or the direction of every band, is used in a clinically valid way.

S3.3 PAIRED RL–BEHAVIOR CONCEPT COMPARISON

Protocol. We compare how the frozen RL and behavior policies use the shared concept dictionary on the same 600-state test cohort. Because the SAE is fitted on the shared representation h, each concept has the same definition for both policies. For each state and policy, we measure a concept's *decision effect* as the change in that policy's decision margin after removing the concept from h and re-running the complete frozen branch, including the behavior-policy adapter. We compare the resulting normalized effect profiles using total-variation distance (TV_e) and the displayed concept sets using Jaccard overlap, separately for policy-agreement and disagreement states. Each statistic is

compared with a permuted-pairing null that pairs the RL and behavior profiles from different states within the same stratum. Confidence intervals use patient-clustered bootstrap resampling.

Results. Across all three cohorts, the two policies are more similar within the same patient state than under permuted pairing, indicating that their decisions draw on related evidence in the shared concept space (Table S24). However, their displayed concept sets become less similar when the policies disagree. The decrease in raw card overlap is clear on NIV (-0.075 $[-0.099, -0.051]$) and vasopressin (-0.221 $[-0.251, -0.190]$), while the corresponding change on RRT is small (-0.018 $[-0.041, +0.005]$). Thus, policy disagreement is accompanied by a measurable shift in which shared concepts are emphasized, although the magnitude of this shift varies across cohorts.

Scope. The dictionary is defined on the shared representation h, rather than on the behavior policy's post-adapter representation $h_b = h + g_b(h)$. The analysis therefore compares how shared input concepts affect the two complete decision branches, but does not decompose how the behavior adapter internally recombines those concepts.

Table S24: Paired comparison of RL and behavior-policy concept use. TV_e is the total-variation distance between normalized concept decision-effect profiles, and Card Jaccard measures overlap between the policy-specific concept sets shown to the explainer. Parentheses report the corresponding permuted-pairing null.

Cohort	Policy actions	n	Card Jaccard	TV_e
RRT	Agree	481	0.218 (0.135)	0.382 (0.463)
RRT	Disagree	119	0.200 (0.165)	0.377 (0.418)
NIV	Agree	524	0.275 (0.065)	0.385 (0.742)
NIV	Disagree	76	0.199 (0.065)	0.384 (0.680)
Vasopressin	Agree	394	0.871 (0.420)	0.312 (0.646)
Vasopressin	Disagree	206	0.650 (0.534)	0.227 (0.355)

S4 Extended Experimental Setup, Results, and Discussion

This section provides the action-level comparisons, paired example, and diagnostic analyses summarized in the main paper, followed by the detailed baseline configurations.

S4.1 Detailed Action Recovery and Policy Comparison

On the paired RRT test cohort (600 states, 263 patients), shared PPO at iteration 2604 reaches RL/behavior action-recovery accuracy of 0.755/0.935, against 0.742/0.897 for its Concept SFT initialization and 0.697/0.873 for State SFT. The w/o behavior policy explainer (Table 2) reaches 0.773 RL accuracy on the same states, 1.8 points higher (patient-cluster bootstrap interval $[-1.5, +5.3]$); because the two runs do not share a rollout budget (Section S2), this bounds the cost of serving both policies rather than isolating it. On the 119 disagreement states the judge recovers each policy's action far more often from that policy's explanation than from the other policy's explanation of the same state (0.706 versus 0.101 for the RL action, 0.899 versus 0.193 for the behavior action; Figure S2; all cohorts in Table S10; outcome groups and controls in Section S3.1).

Head-to-head explanation preference. Main Figure 3 asks which of a state's two stored explanations makes more sense. For each of the 600 states, Claude Opus 5 (effort low) saw the patient observation and the RL and behavior explanations as "Explanation 1" and "Explanation 2" with no policy tag or action (none of the 1,200 explanations contains policy wording) and returned the probability that Explanation 1 makes more sense on clinical coherence and grounding in the observed data; each state was judged in both orders, and the score is the order-averaged probability that the RL explanation makes more sense. Overall the score is 0.503 [0.496, 0.511] (patient-cluster 95% interval): 238 states favor the RL explanation, 273 the behavior explanation, and 89 are tied (sign test $p = 0.13$). Opus's clearest preference is for the RL explanation on the 39 states where the RL policy selects IHD and the behavior policy No-RRT (0.611 [0.593, 0.629]; 36 states against 1);

these patients have median BUN 75 mg/dL, creatinine 4.9 mg/dL, and 6-hour urine output 15 mL. The preference for the behavior explanation when both policies select CRRT (0.461 [0.452, 0.471]) tracks length: across all states behavior explanations are on average 92 characters longer, and at equal length the score is 0.523 [0.499, 0.548] on these states and 0.530 [0.522, 0.539] over all states, a mild lean toward the RL explanation. A preference for one policy's explanation measures how plausible the text reads to one judge; it is not evidence that the corresponding action is clinically better.

Second judge on RRT. GPT-6 Astra (through the Codex CLI, reasoning effort low) judged the same 1,200 tasks, byte-identical to those Opus 5 saw, including which explanation appears first (main Figure 3b). It has no overall preference either (0.493 [0.478, 0.508]; 303 states against 279, sign test $p = 0.34$) and leans toward the RL explanation on the 119 disagreement states (0.536 [0.500, 0.572]; 70 against 49, $p = 0.066$). Where it leans differs from Opus: it prefers the RL explanation where the RL policy stops CRRT that the behavior policy continues (0.660 [0.611, 0.708]; 28 against 6), not where the RL policy starts IHD (0.452 [0.385, 0.518]; 18 against 21). The two judges' per-state scores are nearly uncorrelated (Pearson 0.15 over all states, 0.05 on disagreement states), although GPT-6 Astra is far more consistent across the two orders (order correlation 0.96, against 0.64 for Opus). Preferences for single directions are therefore judge-specific; the absence of an overall preference and the mild lean toward the RL explanation on disagreement states hold under both judges.

Head-to-head preference on NIV and vasopressin. Figure S1 repeats the comparison on the NIV and vasopressin test states with each cohort's TRACE explanations of record, changing only the decision named in the prompt. On NIV the judge has no preference: 0.505 [0.499, 0.510] over 600 states (248 states against 251; sign test $p = 0.93$), and none on the 76 disagreement states (0.494 [0.478, 0.509]). On vasopressin it leans mildly toward the RL explanation: 0.525 [0.517, 0.534] (333 states against 170), most on the 206 disagreement states, where the RL policy initiates vasopressin and the behavior policy does not (0.544 [0.527, 0.560]; 127 against 50). Neither result is a length effect: the two explanations differ by 11 (NIV) and 27 (vasopressin) characters on average, and the scores at equal length are 0.506 [0.500, 0.512] and 0.531 [0.523, 0.539].

Table S15 shows one of the 39 RRT states where the RL policy selects IHD and the behavior policy No-RRT. Both explainers describe the same anuric patient with creatinine 4.7 mg/dL and cite an overlapping set of measurements, but the RL explanation closes on efficient short-interval correction and the behavior explanation on continued observation.

S4.2 Uncertain Concepts on the Explainer's Page

Concept selection is status-blind: unnamed SAE features retain their ranked positions rather than being replaced by lower-ranked named features, although the explainer is not allowed to cite them as evidence. Uncertain cards are common on RRT, present on a smaller fraction of NIV pages, and rare on OVISS. Recovery varies across pages with different numbers of uncertain cards, but these strata are strongly confounded by the underlying action distribution; we therefore do not interpret these differences causally.

S4.3 Baseline Settings

Common settings. Where applicable, all baselines use the same Explanation LLM backbone, patient splits, and evaluation protocol as TRACE. No method receives the target policy's selected action or action probabilities at inference time. Frozen prompting and surrogate-based baselines use the pretrained explainer without task-specific fine-tuning, whereas SFT- and PPO-based baselines use the same teacher-generated training data as TRACE, with concept cards removed for state-only variants. Frozen baselines decode with up to 768 tokens, while fine-tuned baselines use a 384-token generation budget.

Direct LLM. The frozen Explanation LLM generates an explanation directly from the patient state, without concept cards or task-specific training. This baseline measures how much policy-relevant information can be recovered from the clinical state and the pretrained model alone.

Zero-shot CoT. The frozen Explanation LLM receives the patient state and policy-derived concept evidence and is prompted to reason step by step before producing its explanation. No task-specific parameter updates are performed.

MCTS Refinement. Following Belouadah et al. (2025), we use Monte Carlo Tree Search at inference time to iteratively propose and refine candidate explanations from the frozen Explanation LLM. To match our action-blind setting, the target policy action is withheld from both generation and search, although the original method conditions on the action. Candidate explanations use a 384-token budget and critiques use up to 768 tokens.

SFT. We fine-tune the Explanation LLM on class-balanced teacher-generated explanations using the patient state alone, with concept cards removed and no PPO optimization. The teacher may use the private target action when constructing the training target, but the student does not receive the action.

CoT + SFT. This baseline uses the same state-only inputs and teacher data as SFT, but the supervised targets are written as step-by-step clinical reasoning. It therefore tests whether explicit chain-of-thought-style supervision improves policy recovery without concept conditioning or reinforcement learning.

PPO + BT reward. This baseline uses the same Concept-SFT initialization and concept-conditioned inputs as TRACE, but replaces TRACE's recovery-based reward with a Bradley–Terry (BT) preference reward model. The BT model is trained on pairs in which the preferred response is the teacher explanation for the target action and the rejected response is a same-policy-tag teacher explanation associated with a different action. PPO then optimizes the Concept-SFT explainer against this reward, and checkpoints are selected using the development BT reward. For vasopressin, the BT reward is trained on RL-target pairs only because the behavior-policy target provides insufficient contrasting actions.

BC-LLM. Following Feng et al. (2025), we adapt the Bayesian Concept Bottleneck Model with LLM priors to predict each frozen policy's action. The method discovers a sparse set of clinical concepts from the patient state and fits a multinomial logistic regression from these concepts to the policy action. At inference time, the frozen Explanation LLM verbalizes the patient-specific concept annotations, ordered by their contribution to the surrogate prediction; neither the surrogate's predicted action nor its coefficients are exposed to the explainer. Separate bottlenecks are fit for the learned and behavior policies.

Policy Distillation + LLM. Following Zhang et al. (2025), each frozen policy is distilled into a decision-tree surrogate. For each patient state, the corresponding tree path is provided to the frozen Explanation LLM, which verbalizes the decision-relevant rules along that path. The original policy action and action scores are not directly provided to the explainer.

RLAIF with CNF rewards. Following Yang et al., we train a continuous normalizing flow (CNF) reward model from feedback produced by the frozen Qwen3-8B proxy reader over Concept-SFT explanations, and optimize the explainer with PPO from the same Concept-SFT initialization as our w/o-PPO ablation. The reward conditions on the explainer input and generated explanation, with a state-blind variant for NIV. PPO is trained on RL-target examples, while behavior-policy results use the same resulting checkpoint under the behavior-policy tag. We evaluate iterations 744, 417, and 607 for RRT, NIV, and vasopressin, respectively.

S4.4 Leakage Audit

Criteria. The *action-leak* expression is the teacher generator's cohort lexicon, the teacher data's acceptance rule; besides reader label letters ("option A"), it matches terms such as RRT, dialysis, IHD, and CRRT (RRT), intubation, mechanical ventilation, BiPAP, and CPAP (NIV), and start/withhold/continue/stop verbs on vasopressin or a pressor (vasopressin). The *policy-identity* expression matches wording such as RL, BCQ, behavior policy, imitation, and learned policy; numerical *grounding* is the tolerance check of Section S2.

Leakage across arms. Table S25 audits the natural-test explanations of every arm in the main table. Every fine-tuned arm (SFT, CoT+SFT, PPO+BT, RLAIF, the w/o PPO and w/o concepts ablations, and TRACE) has zero action-leak and policy-identity rates on both targets of all three cohorts, so the fine-tuned arms differ in what the judge recovers, not in whether the action is named. Only frozen arms leak: Direct LLM names the action on 8% (RRT), 98% (NIV), and 96% (vasopressin) of RL-target explanations and identifies the policy in every one; zero-shot CoT and MCTS refinement name the action on 10–25% of NIV and vasopressin explanations, BC-LLM on 38% of vasopressin explanations, and the policy-distillation arm carries the tree's action-labelled leaves. Leakage can therefore only inflate the frozen baselines, none of which exceeds TRACE; no rows are removed for leakage.

Leakage before, during, and after PPO. Table S26 follows the leak indicator through the three shared PPO runs, on every logged rollout batch and every development-set evaluation. The rate is zero on the first rollout batch (sampled from the Concept-SFT initialization before any update) and on the natural test set before and after PPO; on the development set it is zero at every evaluation except one (0.002 at iteration 1,400 of the NIV continuation round). The only other non-zero entries are isolated training batches in which a single rollout of six matched the lexicon, removed by the penalty within the next update; PPO's gain over its SFT initialization is measured at zero test leakage.

A length cue that PPO introduced. Under PPO, the number of measurements an explanation quotes came to co-vary with the decision: TRACE's RRT behavior explanations cite 6.6 numbers on average for No-RRT and 4.5 for CRRT, against 6.2 and 5.9 for the Concept-SFT initialization; the same widening appears on NIV but not on vasopressin. A zero-shot judge has no labelled set from which to learn such a cue. The shuffled-explanation control (recovery 0.085–0.307 across the five targets with a shuffled arm; Table S20) shows that the judge follows the explanation it is given, but it does not bound reliance on form, because a shuffled explanation carries the form as well as the evidence of an explanation of a different action.

What the lexical audit does not exclude. The explainer never receives the action, so any information about the decision in its text comes from the state, the policy tag, and the policy-selected concepts. Those concepts are selected by their effect on the policy's chosen action and therefore carry the policy's evidence for that action by construction; conveying that evidence is the purpose of the explanation. The teacher sees the action only to write SFT targets; test explanations are generated without it. The audit shows that no fine-tuned arm's text matches the action or policy-identity lexicon; it cannot exclude that the decision is also encoded in style, such as wording or length, of which the length cue above is one channel. Two further results bear on this channel without bounding it. Replacing the policy-selected cards with size-matched unselected concepts lowers RL recovery by 0.158, 0.152, and 0.075 on RRT, NIV, and vasopressin (Table S14), so recoverability depends on the supplied evidence rather than on a fixed stylistic code; and on RRT, removing the concepts an RL explanation cites changes the RL decision more often than removing equally many uncited page cards (Section S3.2). Neither result shows that the recovered signal is carried by the evidence rather than by phrasing, and much of it is not: with the phrasing that states what the evidence implies for the decision removed and every cited concept and measurement kept, recovery falls substantially on RRT and NIV and slightly on vasopressin (Section S2).

S4.5 Recovery of Infrequent Actions

The matched 600-state RRT test subset contains 359 No-RRT, 39 IHD, and 202 CRRT decisions of the frozen RL policy, so IHD accounts for only 6.5% of targets. Its recall shows whether an explanation conveys evidence for a rare action, and its precision whether that comes from over-predicting it. Table S3 gives the per-action comparison.

Shared PPO raises IHD recall from 0.846 under Concept SFT to 0.949 at unchanged precision (0.303 versus 0.298); the judge predicts IHD for 124 states against 39 IHD targets. CRRT precision improves from 0.817 to 0.879 and F1 from 0.794 to 0.814, while recall changes from 0.772 to 0.757 (Section S3.1). The rare action separates the arms by what the explainer was trained to optimize. The three fine-tuned arms trained on recovery-bearing supervision (SFT, CoT + SFT, TRACE) reach IHD recall 0.692 to 0.949, while the five arms that decode from the frozen explainer stay at or below

Table S25: Regex leakage audit of natural-test explanations ($n = 600$ per cohort and target policy unless noted); each cell gives RL/behavior. "Leak" is the cohort action-word regex hit rate (%), "Ident." the policy-identity regex hit rate (%), and "Words" the mean explanation length. "–" marks a target with no generation on the 600 test states (the vasopressin BC-LLM behavior cell of the main table uses a separate clinician-action subset). The RRT Direct LLM row audits the 768-token decoding used by the Direct LLM arms of the other cohorts. † the NIV RLAIF checkpoint produces empty or near-empty explanations, so its zero rates carry no information.

	RRT			NIV			OVISS		
Arm	Leak	Ident.	Words	Leak	Ident.	Words	Leak	Ident.	Words
Direct LLM	8.3/57.5	100.0/100.0	508/442	97.5/84.2	100.0/100.0	510/447	95.5/88.0	100.0/100.0	466/429
Zero-shot CoT	0.0/0.0	0.0/0.0	104/101	25.0/16.3	0.0/0.0	121/116	19.7/19.7	0.0/0.0	101/99
MCTS Refinement	0.2/0.2	0.2/0.0	147/144	17.2/15.3	1.8/0.0	166/164	10.3/12.2	0.0/0.0	139/137
BC-LLM	0.0/0.0	5.7/0.0	552/534	9.0/1.7	0.0/0.0	485/507	38.3/–	0.0/–	434/–
Policy distill. + LLM	0.8/0.5	53.2/5.3	332/291	0.8/7.7	0.0/6.8	241/318	6.0/48.3	0.0/99.8	390/268
SFT	0.0/0.0	0.0/0.0	69/69	0.0/0.0	0.0/0.0	75/75	0.0/0.0	0.0/0.0	71/71
CoT + SFT	0.0/0.0	0.0/0.0	70/70	0.0/0.0	0.0/0.0	74/74	0.0/0.0	0.0/0.0	71/72
PPO + BT	0.0/0.0	0.0/0.0	89/88	0.0/0.0	0.0/0.0	73/73	0.0/0.0	0.0/0.0	90/93
RLAIF (CNF rewards)	0.0/0.0	0.0/0.0	233/255	0.0/0.0	0.0/0.0	$0^{\dagger}/0^{\dagger}$	0.0/0.0	0.0/0.0	89/92
w/o PPO (Concept SFT)	0.0/0.0	0.0/0.0	79/82	0.0/0.0	0.0/0.0	75/75	0.0/0.0	0.0/0.0	91/94
w/o concepts	0.0/0.0	0.0/0.0	68/70	0.0/0.0	0.0/0.0	73/74	0.0/0.0	0.0/0.0	68/68
TRACE (Ours)	0.0/0.0	0.0/0.0	80/89	0.0/0.0	0.0/0.0	75/75	0.0/0.0	0.0/0.0	91/96

Table S26: Action-leak rate (hit rate of the cohort action-leak expression, %) of TRACE explanations before, during and after PPO. "It. 1" is the first rollout batch, sampled from the Concept-SFT initialization before any policy update; "ckpt" is the mean over the ± 25 iterations around the selected checkpoint; "max" is the largest single-iteration rollout-batch rate over the whole run; "dev@ckpt" is the held-out dev-set rate at the selected checkpoint; "test pre/post" are the natural-test rates of the Concept-SFT parent (w/o PPO) and of TRACE. NIV rows cover the first 1,390 iterations, with "ckpt" and "dev@ckpt" at that round's selected iteration 1225; over the continuation to 2,780 iterations development leakage (as a fraction) was 0.000 at every evaluation except 0.002 at iteration 1,400.

Cohort	Target	It. 1	ckpt	max (it.)	dev@ckpt (n)	test pre	test post
RRT	RL	0.0	0.0	0.0 (1)	0.0 (623)	0.0	0.0
	Behavior	0.0	0.0	0.0 (1)	0.0 (623)	0.0	0.0
NIV	RL	0.0	0.0	0.0 (1)	0.0 (289)	0.0	0.0
	Behavior	0.0	0.0	16.7 (494)	0.0 (289)	0.0	0.0
OVISS	RL	0.0	0.0	16.7 (118)	0.0 (359)	0.0	0.0
	Behavior	0.0	0.0	16.7 (176)	0.0 (359)	0.0	0.0

0.205 (Direct LLM 0.026, Zero-shot CoT 0.000, MCTS Refinement 0.026, Policy Distillation + LLM 0.077, BC-LLM 0.205). PPO + BT, which starts from the same Concept-SFT checkpoint as TRACE and differs only in the reward (and its selection metric), falls to IHD recall 0.077, CRRT recall 0.188, and 0.478 overall, below the 0.598 majority rate. Within the frozen group BC-LLM's ranked concepts help relative to Direct LLM: against its like-for-like 768-token run, IHD F1 rises from 0.000 to 0.157 and accuracy from 0.452 to 0.487 (from 0.045 and 0.437 for the 384-token run in Table S3), with IHD precision 0.127 because the surrogate itself recovers IHD with recall 0.10.

Why BC-LLM stays near the majority rate. On RRT the bottlenecks are not the weak link: they predict the RL and behavior actions with accuracy 0.685 and 0.768 (majority 0.598 and 0.683), and the judge agrees with them on 0.575 and 0.642 of states; the residual gap comes from the RL surrogate (CRRT recall 0.47, IHD 0.10), three of whose eight concepts need evidence the state lacks. On NIV the bottleneck itself fails (RL 0.522, majority 0.497): no concept measures gas exchange, the physiology-only proposer excludes the variables a distilled tree splits on first (elapsed time on NIV, oxygen flow), and the behavior bottleneck predicts Continue on 597 of 600 states.

Why the Bradley–Terry reward degrades the explainer. PPO + BT differs from TRACE only in the reward and in selecting its checkpoint on the development BT reward, which rises from 10.454 (iteration 372) to 11.226 (2975), yet the judge recovers the RL action on only 0.478 of RRT states, below the 0.598 majority rate. The explainer applies one hedged register of observation and reassessment to 83.6%, 61.5%, and 93.1% of No-RRT, IHD, and CRRT states and drops the wording that separates continuous from intermittent therapy, used by its SFT parent in 78.2% of CRRT states. This motivates a reward defined on policy recovery rather than on preference between explanations.

NIV and vasopressin per-action recovery. Tables S4 and S5 repeat the per-action breakdown for the other two cohorts. On NIV, PPO leans toward the rare action: intubation is recovered on 4 of 6 states rather than 3 of 6, with the judge assigning it to 27 rather than 17 states; with six such states these cells should be read as counts only. On vasopressin, where initiation is 34% of RL targets, PPO trades initiation recall for precision (recall 0.723 to 0.602, precision 0.703 to 0.785), and the w/o PPO and w/o concepts arms bracket TRACE in accuracy (0.800 and 0.818 against 0.807) and exceed it in initiation F1 (0.713 and 0.725 against 0.681).

S4.6 Judge Robustness

The main-table numbers come from one judge, Claude Opus 5. We re-score the same frozen matched explanations of five arms per cohort (TRACE, w/o PPO, w/o concepts, state-only SFT, and one external baseline per cohort: BC-LLM on RRT, PPO+BT on NIV, and RLAIF on vasopressin) with Claude Sonnet 5 under the identical blinded protocol, and measure per-row agreement as Cohen's κ between Sonnet's and Opus 5's action predictions (Table S27).

Per-row agreement on TRACE's explanations is $\kappa = 0.885/0.839$ on RRT (RL/behavior), $0.883/0.913$ on NIV, and 0.904 on the vasopressin RL target; on the one-class vasopressin behavior target κ is degenerate (0.253). Against the external baseline and state-only SFT, TRACE ranks first under Sonnet on five of the six cohort–target panels; the exception is the one-class vasopressin behavior target, where Sonnet scores SFT higher (0.943 against 0.918). On the ablations the two judges agree on the direction: TRACE is above w/o PPO and w/o concepts on RRT and NIV, and removing concepts scores higher than TRACE on vasopressin under Sonnet as under Opus (RL 0.813 against 0.787; behavior 0.978 against 0.918). The PPO gain over the Concept-SFT initialization (w/o PPO) is larger under Sonnet: $+0.105$ $[+0.059, +0.153]$ on RRT RL (Opus: $+0.013$) and $+0.307$ $[+0.253, +0.356]$ on NIV RL (Opus: $+0.137$), because Sonnet reads the Concept-SFT explanations much lower than Opus does (0.605 against 0.742 on RRT, 0.513 against 0.705 on NIV) and TRACE's by less (0.710 against 0.755; 0.820 against 0.842). Outside the one-class vasopressin behavior target, TRACE therefore ranks first under both judges against the two re-scored baselines (state-only SFT and the cohort's external baseline), although under Sonnet its RRT lead over state-only SFT is small (0.710 against 0.687 on RL, 0.868 against 0.852 on behavior); the other main-table baselines were not re-scored. On vasopressin, the w/o concepts result and the one-class behavior target qualify the column under both judges.

Table S27: Judge-swap robustness. The same frozen matched explanations are scored by the published Claude Opus 5 judge and by Claude Sonnet 5 under the identical blinded protocol (state, policy tag, explanation, per-row label permutation; no cards, no action). Acc. is action-recovery accuracy; κ is Cohen's kappa between Sonnet 5's and Opus 5's per-row action predictions. SFT is the state-only SFT baseline and w/o PPO the Concept-SFT initialization. The external baseline per cohort is BC-LLM (RRT), PPO+BT (NIV), and RLAIF with CNF rewards (vasopressin). The vasopressin behavior target is one-class, so its κ values are degenerate.

		RL target			Behavior target		
Cohort	Arm	Opus 5 Acc.	Sonnet 5 Acc.	κ	Opus 5 Acc.	Sonnet 5 Acc.	κ
RRT	**TRACE (Ours)**	0.755	0.710	0.885	0.935	0.868	0.839
	w/o PPO	0.742	0.605	0.522	0.897	0.648	0.289
	SFT	0.697	0.687	0.956	0.873	0.852	0.925
	w/o concepts	0.612	0.585	0.840	0.838	0.783	0.734
	BC-LLM	0.487	0.503	0.539	0.608	0.500	0.336
NIV	**TRACE (Ours)**	0.842	0.820	0.883	0.892	0.873	0.913
	w/o PPO	0.705	0.513	0.371	0.788	0.648	0.374
	SFT	0.545	0.500	0.402	0.578	0.563	0.454
	w/o concepts	0.532	0.425	0.365	0.637	0.572	0.382
	PPO+BT	0.618	0.388	0.134	0.705	0.533	0.110
Vasopressin	**TRACE (Ours)**	0.807	0.787	0.904	0.975	0.918	0.253
	w/o PPO	0.800	0.773	0.700	0.917	0.695	0.284
	SFT	0.765	0.755	0.911	0.957	0.943	0.474
	w/o concepts	0.818	0.813	0.981	0.983	0.978	0.690
	RLAIF (CNF)	0.773	0.738	0.794	0.900	0.677	0.331

S5 NIV Teacher Data, SFT, and Shared PPO

Frozen policies and cards. On the jointly trained three-head BCQ of Section S1, the RL target is the masked-greedy action and the behavior target the clinician-head argmax. Each state shows eight policy-specific cards from the validation-selected SAE (width 512, native Top-$K = 128$;

Teacher data. The paired teacher set holds 4,458 subject-train state–policy pairs (8,916 records): 2,000 discontinue and 2,000 continue states sampled by deployed BCQ action and all 157 real and 301 synthetic intubation states (flagged local perturbations of real ones that stay intubation decisions under the frozen support rule). The policies disagree on 536 pairs (12.0%). The teacher is GPT-5.6 Terra under the RRT contract; all records of both views passed every deterministic gate. A subject-hash holdout of 104 development patients (578 records) leaves 8,338 training records per view.

SFT. Both arms start from the locked Llama-3.1-8B-Instruct snapshot: three epochs over the 8,338 records (2,085 optimizer steps), effective batch 12, learning rate 2×10^{-5}, seed 42, assistant-token loss. State-only sequences span 472–611 tokens (final loss 0.411, 1.1 hours on one H200); card-aware sequences span 1,721–5,414 tokens (mean 3,987; 4.1 supported band-resolved axes per card), untruncated at a 6,144-token maximum (final loss 0.522, 5.6 hours).

Final test protocol. Both arms are scored under the matched Opus 5 protocol (four arms: matched, cross-policy, state-only, shuffled same-tag different-action; six label permutations; patient-cluster bootstrap) on the frozen 600-state natural test cohort (194 patients, 1,200 policy-tagged rows; RL actions 296/298/6 and behavior actions 222/374/4 for discontinue/continue/intubate; 76 disagreement states). All 2,400 generations are nonempty and EOS-complete with a 384-token limit, and none matches the prespecified action-leakage expression. Macro-F1 averages the classes present.

State SFT and Concept SFT. State SFT (SFT row) reaches matched accuracy 0.545 (RL) and 0.578 (behavior), macro-F1 0.420 and 0.403, against state-only 0.513 and 0.625 (matched minus state-only $+0.032$ $[-0.032, +0.096]$ and -0.047 $[-0.104, +0.011]$): explanations from the state text alone add nothing over that text. Concept SFT (w/o-PPO row), which initializes shared PPO,

reaches 0.705 and 0.788 against state-only 0.508 and 0.625 ($+0.197$ $[+0.143, +0.253]$ and $+0.163$ $[+0.113, +0.212]$); its policy specificity is confined to the 76 disagreement states (Table S28).

Shared PPO (TRACE). The shared explainer continues from the Concept SFT checkpoint under the explanation reward of Section S2, with batch 12, 384 new tokens, a 6,144-token prompt cap, card-use weight 0.3, and the frozen Qwen3-8B reader. The 578 development records are evaluated every 175 iterations, checkpoints are written every 25, and selection maximizes development recovery with early stopping disabled. Training runs in two rounds under the same recipe: 1,390 iterations (two epochs) at learning rate 2×10^{-5} over several allocations, then, from that round's final checkpoint, to 2,780 iterations (four epochs) at 1×10^{-5}. Development accuracy peaks at 0.867 (macro-F1 0.804) at iteration 2275; this second-round checkpoint, selected over the first round's best (0.777) on development accuracy, is the one reported. Action leakage is 0.000 at every evaluation but one (0.002 at iteration 1400), and card use and numeric grounding stay at or above 0.997. The comparison arms were trained for fewer iterations: w/o concepts and PPO+BT for 1,390 and the RL-only w/o behavior policy arm for 695 (two passes over their rows), and RLAIF for 417 of a planned 695. At the first round's budget of 1,390 iterations, that round's development-selected checkpoint (iteration 1,225, development accuracy 0.777) reaches 0.828 (RL) and 0.832 (behavior) on the test cohort, above every NIV baseline and ablation except w/o behavior policy (0.852).

Table S28: NIV Concept SFT and the TRACE checkpoint of record (iteration 2275, selected on development recovery) under the Opus 5 protocol. Matched, cross, shuffled, and macro-F1 use all 600 states per policy; the last column is matched-minus-cross accuracy on the 76 disagreement states (95% patient-cluster bootstrap intervals). Each row carries the controls from its own judge run.

Checkpoint	Policy	Matched	Cross	Shuffled	Macro-F1	Gap on disagreement
Concept SFT	RL	0.705	0.713	0.295	0.557	$+0.197$ [0.093, 0.316]
TRACE PPO (2275)	RL	0.842	0.767	0.125	0.653	$+0.513$ [0.386, 0.632]
Concept SFT	Behavior	0.788	0.817	0.333	0.603	$+0.171$ [0.081, 0.263]
TRACE PPO (2275)	Behavior	0.892	0.840	0.190	0.662	$+0.526$ [0.403, 0.644]

Within-state paired policy recovery. On the 76 NIV disagreement states, the joint measure of Section S3.1 (both policies' actions recovered under their own tags) rises from 16/76 (0.211; 95% interval [0.100, 0.329]) under Concept SFT to 39/76 (0.513; [0.377, 0.640]) under shared PPO, a paired difference of $+0.303$ [0.164, 0.443]. The rest are 2 RL-only, 57 behavior-only, and 1 unrecovered under Concept SFT, and all 37 behavior-only (the majority-action direction) after PPO.

Reading the NIV shared-PPO result. Shared PPO adds 13.7 (RL) and 10.4 (behavior) accuracy points over Concept SFT and $+0.293$ [0.238, 0.349] and $+0.267$ [0.219, 0.314] over the state-only judge; on disagreement states the matched-minus-cross gap reaches $+0.513$ and $+0.526$, and cross-arm predictions follow the *other* policy's action on all 76 RL disagreement states. Macro-F1 is dominated by intubation, the target on only six RL and four behavior test states (RL intubation precision 0.148 at recall 0.667), so macro-F1 differences of this size are not changes in explanation quality. The state-only judge already answers the behavior target correctly on all 76 disagreement states, so the RL target is the informative one there (0.211 state-only against 0.513 matched; Table S20).

S6 Vasopressin Teacher Data, SFT, and Shared PPO

Frozen policies and cards. The two targets are defined as on NIV, from the jointly trained three-head BCQ (Section S1.1). Cards come from the validation-selected whitened SAE (width 64, $K = 6$). Every state has six active concepts; each page shows four, chosen for its policy by the RRT and NIV ranking and an axis-diverse selection, with decoder directions mapped back from the whitened metric to the policy representation before any effect is computed.

Teacher data. The paired teacher set holds 4,000 state–policy pairs (8,000 records), 2,000 per deployed RL action, drawn without replacement (seed 42) from training-split decision states of 1,370 patients; no synthetic states are added. The estimated behavior target initiates on 2 of these

states, so the policies disagree on 1,998 pairs (50.0%). The teacher is GPT-5.6 Terra under the RRT and NIV contract with four cards per record; numbers shown on the cards count as grounded and a named yes/no card axis counts as cited evidence. All 8,000 records passed every deterministic gate, from 7,949 unique card-aware teacher calls, and a subject-hash holdout of 128 development patients (718 records) leaves 7,282 training records per view.

SFT. Both arms are fully fine-tuned from the locked Llama-3.1-8B-Instruct snapshot for three epochs over the 7,282 records (effective batch 12, learning rate 2×10^{-5}, seed 42); the 4,096-token maximum length truncates no record (the longest card-aware record is about 3,000 tokens).

Shared PPO (TRACE). The shared explainer continues from the Concept SFT checkpoint under the explanation reward and optimization settings of Section S2, with four cards per record, two label permutations, and a 5,632-token prompt cap. It runs 1,214 iterations (two passes over the 7,282 paired training records), and development recovery selects the final iteration.

Final test protocol. All arms are scored under the matched Opus 5 protocol (the four arms of NIV; two label permutations; patient-cluster bootstrap) on a frozen cohort of 600 decision states drawn uniformly without replacement (seed 42) from the 5,076 test-split decision states: 166 of the 311 test patients, 1,200 policy-tagged rows, and no state or patient shared with the teacher data. The trained explainers decode greedily with a 384-token limit, and all their explanations are EOS-complete. The RL policy initiates vasopressin on 206 of the 600 states and the estimated behavior policy on none; the policies thus disagree on exactly these 206 states, and the behavior target is a single class: its accuracy is the rate at which the judge answers no vasopressin (its macro-F1 is the F1 of that class), so a constant answer would score 1.000, and we do not read the behavior column as evidence of explanation quality on this cohort (Section S1). With no behavior explanation of the other action to transplant, the shuffled arm is scored on the 600 RL rows only.