DYNAMIC CAUSAL TREE: INTERVENTIONAL CREDIT ASSIGNMENT FOR LLM AGENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning improves multi-turn decision making in LLM agents, but trajectory-level rewards obscure the contribution of individual decisions. Failed trajectories can suppress credit for useful decisions, while successful ones can reinforce harmful decisions. We formalize this mismatch through same-history interventions, separate causal relevance from policy-update credit, and bound how credit error distorts the expected policy update. We propose *Dynamic Causal Tree* (DCT), which selectively constructs shared-prefix branches to approximate exhaustive intervention comparisons. The resulting tree supports a dual readout: pairwise branch contrasts estimate causal relevance, while policy-weighted aggregation yields signed credit. Experiments on multi-hop QA across model scales show consistent gains over trajectory-level and tree-based RL baselines, while independent intervention tests support more faithful decision attribution.

1 INTRODUCTION

Reinforcement learning (RL) can improve multi-turn decision making in language-model agents. In these tasks, an agent repeatedly acts on partial observations, invokes tools, and adapts over long interaction horizons (Xi et al., 2026a; Feng et al., 2026). Training, however, commonly relies on a single trajectory-level outcome (Feng et al., 2025; Xie et al., 2026). Such an outcome reveals whether the interaction succeeded, but not which intermediate decisions helped or harmed it.

Prior work has highlighted mismatches between final outcomes and intermediate-step quality (Guo et al., 2025; Yang et al., 2026). We distinguish two complementary attribution failures. A failed trajectory may contain decisions that improve the chance of success but are obscured by later failure; we call this *underthinking*. A successful trajectory may instead contain decisions that reduce the chance of success but are reinforced with the whole trajectory; we call this *overthinking*. These are attribution failures rather than properties of trajectory length: factual outcomes record what happened, whereas credit assignment asks what would have changed under a different decision.

Existing approaches address this ambiguity from complementary directions. Process rewards provide dense supervision over intermediate behavior(Cheng et al., 2025; Zhang et al., 2025b; Xie et al., 2026); critics estimate downstream returns to support lower-variance updates(Huang et al., 2026; Liu et al., 2026); and rollout-based estimators ground credit in sampled future outcomes (Kazemnejad et al., 2025; Guo et al., 2025). These signals improve optimization, but finer granularity and better outcome prediction do not establish causal attribution without same-history decision comparisons under a common downstream policy.

We therefore study agentic credit assignment through same-history interventions: hold the interaction history fixed, replace the current decision, and let the remaining interaction proceed under the same policy. This view separates two requirements that are often conflated. Causal relevance asks whether any feasible replacement can change the expected outcome, whereas optimization consistency asks whether credit preserves the update induced by the original task objective. Neither requirement subsumes the other: aggregation can hide outcome-changing alternatives, while an optimization-consistent signal can still misattribute individual decisions. We formalize this distinction, derive optimization-compatible interventional credit, and bound the update deviation caused by credit error.

Exhaustive intervention evaluation is nevertheless impractical: every visited history may admit many alternative decisions, each requiring a downstream rollout. To make these comparisons tractable,

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.

we propose *Dynamic Causal Tree* (DCT). DCT reuses shared prefixes to selectively approximate same-history interventions, retaining cross-branch outcomes rather than treating a selected path as the learning target. Tree expansion thus turns sparse terminal outcomes into decision-level causal credit. Our contributions are three-fold:

- We identify and causally formalize an outcome–credit mismatch in multi-turn RL, where success can reinforce harmful decisions and failure can suppress useful ones.
- We establish causal relevance and optimization consistency as distinct requirements, and derive a bound linking credit-estimation error to policy-gradient deviation.
- We propose *Dynamic Causal Tree* (DCT), which uses shared-prefix branching to approximate same-history intervention comparisons and estimate turn-level credit for policy optimization.

2 Related Work

Agentic reinforcement learning trains LLMs for sequential reasoning and interaction using task feedback, spanning search, web exploration, and model orchestration (Chen et al., 2025; Li et al., 2025b; Zhang et al., 2025a). ReSearch (Chen et al., 2025) learns when and how to retrieve, while Router-R1 (Zhang et al., 2025a) optimizes model invocation and aggregation. Despite performance gains, trajectory-level outcomes alone offer limited attribution, motivating finer-grained credit assignment.

Intermediate feedback comes from process supervision or return-based estimation. Process supervision evaluates task progress (Zhang et al., 2025b; Xi et al., 2026b), providing dense signals whose value depends on objective alignment. Return-based methods estimate local credit from sampled outcomes: VinePPO (Kazemnejad et al., 2025) estimates reasoning-step values through Monte Carlo continuations, while GiGPO (Feng et al., 2025) compares actions at recurring environment states. Yet a zero advantage can mask opposing effects across alternatives.

3 Problem Formulation

Sequential interaction. For task $x \sim \mathcal{D}$, at turn $t = 1, \ldots, T$ the agent samples $A_t \sim \pi_\theta(\cdot \mid H_t)$ given $H_t = (x, O_{\le t}, A_{<t})$ and observes O_{t+1}, yielding $\tau = (x, O_1, A_1, \ldots, O_T, A_T, O_{T+1})$, where A_t is one complete turn-level decision. With latent state S_t, initial distribution $\rho_1(\cdot \mid x)$, transition kernel P, and observation kernel Ω, the joint interaction factorizes as

$$p_\theta(s_{1:T+1}, o_{1:T+1}, a_{1:T} \mid x) = \rho_1(s_1 \mid x)\Omega(o_1 \mid s_1) \times \prod_{t=1}^{T} \pi_\theta(a_t \mid h_t) P(s_{t+1} \mid s_t, a_t)\Omega(o_{t+1} \mid s_{t+1}). \tag{1}$$

Learning objective. Agentic RL maximizes $J(\theta) = \mathbb{E}[R(\tau)]$ for a binary terminal reward $R(\tau) \in \{0, 1\}$, where the expectation is over the task and interaction above. Since θ appears only through π_θ, its policy gradient weights every decision in a trajectory by the same terminal reward:

$$\nabla_\theta J(\theta) = \mathbb{E}\left[\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(A_t \mid H_t)\, R(\tau)\right] \tag{2}$$

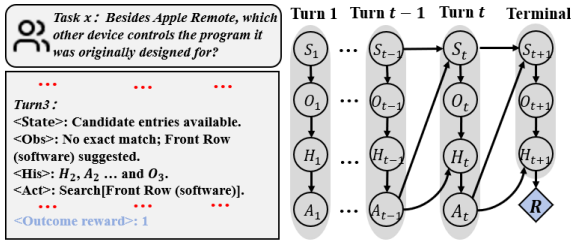

Figure 1: SCM of sequential interaction.

Causal representation. For a fixed task x, the interaction admits an SCM $\mathcal{C} = (\mathbf{V}, \mathbf{F}, P_\mathbf{U})$, illustrated in Figure 1, where $\mathbf{V} = (\mathbf{S}, \mathbf{O}, \mathbf{H}, \mathbf{A}, R)$. With H_t and $R(\tau)$ defined above, the remaining mechanisms are

$$\begin{aligned} S_1 &= f_\rho(x), & O_t &= f_\Omega(S_t), \\ A_t &= f_\pi(H_t; \theta), & S_{t+1} &= f_P(S_t, A_t). \end{aligned} \tag{3}$$

These mechanisms induce Eq. 1; at fixed θ, $\mathrm{do}(A_t = a)$ replaces only the mechanism for A_t.

4 CAUSAL CREDIT ASSIGNMENT

Using same-history interventions, we formalize the distinction between terminal reward and decision-level credit and derive two complementary signals for causal attribution and policy optimization.

4.1 INTERVENTIONAL DECISION CREDIT

At a history $H_t = h$, intervene by setting $A_t = a$, while subsequent decisions follow π_θ. Define

$$\begin{aligned} Q_t^\pi(h,a) &:= \mathbb{E}[R \mid \mathrm{do}(A_t = a), H_t = h], \qquad V_t^\pi(h) := \mathbb{E}_{a' \sim \pi_\theta(\cdot|h)}[Q_t^\pi(h,a')], \\ \Delta_t^\pi(h;a,a') &:= Q_t^\pi(h,a) - Q_t^\pi(h,a'), \qquad\quad A_t^\pi(h,a) := Q_t^\pi(h,a) - V_t^\pi(h). \end{aligned} \tag{4}$$

Thus, $\Delta_t^\pi(h;a,a') > 0$ favors a over a', while $A_t^\pi(h,a)$ averages this contrast over $a' \sim \pi_\theta(\cdot \mid h)$.

Let $s_t = \nabla_\theta \log \pi_\theta(A_t \mid H_t)$. Under the sequential SCM, $Q_t^\pi(h,a) = \mathbb{E}[R \mid H_t = h, A_t = a]$ for actions in the policy support. Since $\mathbb{E}[s_t \mid H_t] = 0$, $V_t^\pi(H_t)$ is a valid baseline, yielding

$$\nabla_\theta J(\theta) = \mathbb{E}\left[\sum_{t=1}^{T} s_t A_t^\pi(H_t, A_t)\right]. \tag{5}$$

Outcome–credit mismatch. The realized outcome and the interventional advantage can disagree:

$$\begin{aligned} &R(\tau) = 0, \quad A_t^\pi(H_t, A_t) > 0 \text{ for some } t \quad \text{(underthinking)}, \\ &R(\tau) = 1, \quad A_t^\pi(H_t, A_t) < 0 \text{ for some } t \quad \text{(overthinking)}. \end{aligned} \tag{6}$$

These mismatches concern attribution. The terminal-reward estimator remains unbiased, whereas estimated decision credit may distort the expected update.

Proposition 1 (Credit Misalignment Bound)**.** *Let $\widehat{A}_t$ be a stop-gradient decision credit and define $\widehat{g}(\theta) := \mathbb{E}[\sum_{t=1}^{T} s_t \widehat{A}_t]$. If $\mathbb{E}\|s_t\|^2 < \infty$ and $\mathbb{E}[\widehat{A}_t^2] < \infty$ for all t, then*

$$\widehat{g}(\theta) - \nabla_\theta J(\theta) = \mathbb{E}\left[\sum_{t=1}^{T} s_t\big(\widehat{A}_t - A_t^\pi(H_t, A_t)\big)\right], \tag{7}$$

$$\|\widehat{g} - \nabla_\theta J\| \le \sum_{t=1}^{T} \sqrt{\mathbb{E}\|s_t\|^2}\, \sqrt{\mathbb{E}\left[\big(\widehat{A}_t - A_t^\pi(H_t, A_t)\big)^2\right]}. \tag{8}$$

History-dependent baselines can loosen the bound without changing the update; causal relevance therefore requires a separate statistic. The detailed proof is provided in Appendix B.2.

4.2 CAUSAL RELEVANCE AND OPTIMIZATION CONSISTENCY

The exact advantage suffices for policy optimization but not causal relevance. $A_t^\pi(h,a) = \mathbb{E}_{a' \sim \pi_\theta(\cdot|h)}[\Delta_t^\pi(h;a,a')]$ is a policy-weighted mean of pairwise effects. Opposing effects can cancel: $A_t^\pi(h,a) = 0$ need not imply zero effect for every feasible replacement.

For a specified feasible intervention set $\mathcal{A}(h)$, define the interventional relevance magnitude

$$\Gamma_t^\pi(h,a) := \sup_{a' \in \mathcal{A}(h) \setminus \{a\}} \left|\Delta_t^\pi(h;a,a')\right|. \tag{9}$$

Thus, $\Gamma_t^\pi(h,a) = 0$ exactly when all feasible replacements preserve the expected outcome. In contrast, $A_t^\pi(h,a) = 0$ with $\Gamma_t^\pi(h,a) > 0$ shows that policy averaging can mask nonzero pairwise effects. Both depend on the downstream policy π_θ, with relevance defined relative to $\mathcal{A}(h)$.

Causal credit mechanisms should preserve relevance evidence and yield valid update weights. These population-level requirements are ideal targets for finite estimators using sampled interventions.

Definition 2 (Causal Credit Requirements)**.** *Suppose a mechanism produces a nonnegative relevance score $\widehat{\Gamma}_t$ and a stop-gradient update credit $\widehat{A}_t$.*

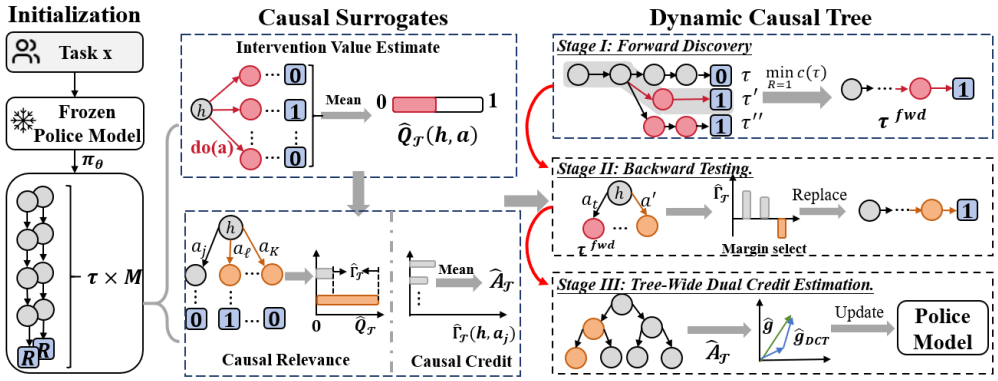

Figure 2: Overview of the proposed Dynamic Causal Tree framework.

(i) Relevance faithfulness. *The relevance vanishes exactly for interventionally invariant decisions:*

$$\widehat{\Gamma}_t(h,a) = 0 \quad \Longleftrightarrow \quad \Gamma_t^\pi(h,a) = 0. \tag{10}$$

(ii) Optimization consistency. *The update credit preserves the original expected policy gradient:*

$$\mathbb{E}\left[\sum_{t=1}^{T} s_t \widehat{A}_t(H_t, A_t)\right] = \nabla_\theta J(\theta). \tag{11}$$

The requirements are complementary: relevance faithfulness concerns outcome-changing interventions, while optimization consistency concerns policy updates. The latter does not fix credit's zero set: adding any history-dependent baseline $b_t(H_t)$ preserves the expected gradient since $\mathbb{E}[s_t b_t(H_t)] = 0$, but can change where credit vanishes. Thus, Γ_t^π and A_t^π serve structural and optimization roles.

Estimating Γ_t^π requires retaining pairwise intervention outcomes before aggregation; estimating A_t^π requires their aggregation with current policy weights. Since exhaustive evaluation is prohibitive over long horizons, we introduce DCT, which reuses shared prefixes to selectively evaluate alternatives.

5 Method

DCT builds shared-prefix intervention trees during training through forward discovery and backward testing (Figure 2). It estimates causal relevance Γ_t^π and signed credit A_t^π from sampled interventional values, then assigns credit to behavior-policy anchor decisions for policy optimization.

5.1 Tree-Structured Intervention Estimation

At each iteration, DCT freezes the behavior policy π_{old} and builds an interaction tree $\mathcal{T} = (\mathcal{H}, \mathcal{E})$ from behavior-policy anchor trajectories. Nodes $h \in \mathcal{H}$ represent histories with replayable pre-decision contexts; edges $(h,a) \in \mathcal{E}$ represent intervened decisions. Sibling branches share a context and continue under π_{old}, realizing same-history interventions (Section 4.1). We assume anchor contexts follow their conditional law given $H_t = h$ and replay independently resamples all post-intervention environment and policy noise.

For an evaluated edge (h,a), let $\{\tau_i^{h,a}\}_{i=1}^{N(h,a)}$ denote its independently sampled continuations. DCT estimates its interventional value by

$$\widehat{Q}_\mathcal{T}(h,a) = \frac{1}{N(h,a)} \sum_{i=1}^{N(h,a)} R\left(\tau_i^{h,a}\right). \tag{12}$$

Let $a_j \sim \pi_{\text{old}}(\cdot \mid h)$, $j = 1, \ldots, K(h)$, be independently sampled outgoing decisions, including the factual decision at an anchor history, with multiplicity retained. DCT estimates relevance when a

distinct alternative is observed and signed credit whenever $K(h) \geq 2$:

$$\begin{aligned} \widehat{\Gamma}_{\mathcal{T}}(h, a_j) &:= \max_{\ell:\, a_\ell \neq a_j} \left| \widehat{Q}_{\mathcal{T}}(h, a_j) - \widehat{Q}_{\mathcal{T}}(h, a_\ell) \right|, \\ \widehat{A}_{\mathcal{T}}(h, a_j) &:= \widehat{Q}_{\mathcal{T}}(h, a_j) - \frac{1}{K(h) - 1} \sum_{\ell \neq j} \widehat{Q}_{\mathcal{T}}(h, a_\ell). \end{aligned} \tag{13}$$

The first statistic retains the largest observed pairwise effect and is therefore relative to the sampled alternatives. The second uses a leave-one-out estimate of the old-policy baseline. If $K(h) < 2$, DCT sets $\widehat{A}_{\mathcal{T}}$ to the factual terminal reward; if no distinct alternative is observed, it leaves $\widehat{\Gamma}_{\mathcal{T}}$ unassessed. DCT thus reads the same tree in two ways: cross-branch variation yields causal relevance, whereas policy aggregation yields signed optimization credit.

5.2 Dynamic Causal Tree

DCT collects up to B additional continuations per anchor. Each expansion samples $a' \sim \pi_{\text{old}}(\cdot \mid h)$, increments $K(h)$ and $N(h, a')$, and updates value and relevance estimates. Repeated draws are retained. Failed anchors enter Stage I; successful anchors enter Stage II using their original paths. Details appear in Appendix C.3.

Stage I: Forward Success-Witness Discovery. Within the forward budget, DCT selects replayable nonterminal histories in the current tree using $\widehat{\Gamma}_{\mathcal{T}}$-guided sampling (Appendix C.3.2). The rule mixes uniform exploration with relevance-proportional sampling and falls back to uniform sampling when all scores vanish. After this allocation, it selects

$$\tau^{\text{fwd}} \in \arg \min_{\tau \in \mathcal{L}(\mathcal{T}):\, R(\tau)=1} c(\tau), \tag{14}$$

where $c(\tau)$ counts interaction turns. This sample-level witness provides the reference for Stage II; its length enters neither reward nor credit formulas. If no success is found, Stage II is skipped.

Stage II: Backward Candidate-Necessity Testing. Using the remaining budget, DCT samples replayable nonterminal histories on the current successful reference using the same rule, scored for reference decisions. At a selected history h_t, let a_t be its reference action. When a distinct alternative exists, define

$$\widehat{m}_{\mathcal{T}}(h_t, a_t) := \widehat{Q}_{\mathcal{T}}(h_t, a_t) - \max_{a' \in \mathcal{A}_{\mathcal{T}}(h_t) \setminus \{a_t\}} \widehat{Q}_{\mathcal{T}}(h_t, a'). \tag{15}$$

Here $\mathcal{A}_{\mathcal{T}}(h_t)$ contains sampled actions. If $\widehat{m}_{\mathcal{T}}(h_t, a_t) \leq 0$ and a best-valued alternative has an observed shorter successful suffix, DCT replaces the reference suffix and refreshes its histories and decisions. All branches are retained; the test is candidate-relative.

Stage III: Tree-Wide Dual Credit Estimation. After forward and backward expansion, DCT consolidates $\widehat{Q}_{\mathcal{T}}$ and $\widehat{\Gamma}_{\mathcal{T}}$ throughout the tree and assigns $\widehat{A}_{\mathcal{T}}$ to the factual decisions on the behavior-policy anchor trajectories.

Let $\{\tau_m^0\}_{m=1}^M$ be the anchor trajectories, and let $s_{m,t}^{\text{old}}$ denote their score-function gradients under π_{old}. DCT forms the update

$$\widehat{g}_{\text{DCT}} = \frac{1}{M} \sum_{m=1}^{M} \sum_{t=1}^{T_m} s_{m,t}^{\text{old}} \, \text{sg}\Big[\widehat{A}_{\mathcal{T}}(H_{m,t}, A_{m,t}) \Big], \tag{16}$$

where sg denotes stop-gradient. Counterfactual branches supply intervention evidence but are not counted as additional on-policy visits. In PPO-style training(Schulman et al., 2017), the same credit is inserted into the standard clipped surrogate.

Relation to the causal objective. DCT targets the original task objective without an auxiliary reward. At the behavior policy and in expectation over anchor contexts, $\widehat{A}_{\mathcal{T}}$ targets $A_t^{\pi_{\text{old}}}$ when $K(h) \geq 2$, with the factual terminal reward as a valid Monte Carlo fallback; $\widehat{\Gamma}_{\mathcal{T}}$ records detected relevance among sampled alternatives. Because Eq. 16 retains the anchor distribution, Proposition 1 links residual credit error to expected-update deviation. The corresponding analysis is given in Appendix B.4.

Table 1: Exact Match (%) on Single-Hop and Multi-Hop QA. Best results are bold.

		Single-Hop QA				Multi-Hop QA				
Model	**Method**	**NQ**	**Trivia**	**PopQA**	**Avg./Δ_{rel}**	**Hotpot**	**2Wiki**	**MuSiQ**	**Bamb.**	**Avg./Δ_{rel}**
Qwen2.5-1.5B	Direct Inference	7.1	22.4	9.9	13.1	5.9	4.3	2.6	8.0	5.2
	Search-o1	10.2	30.9	15.0	18.7	11.6	12.2	3.1	13.0	10.0
	ReAct	9.5	22.1	13.8	15.1	7.3	8.0	1.9	11.2	7.1
	+ GRPO	39.4	51.0	39.7	43.4/Δ_{base}	14.6	24.4	2.2	4.0	11.3/Δ_{base}
	+ GSPO	36.8	48.9	37.3	41.0/-5.5%	15.8	23.7	2.5	4.8	11.7/$+3.5\%$
	+ Tree-GRPO	43.6	**57.3**	**41.6**	**47.5**/$+9.5\%$	29.5	26.8	6.6	13.6	19.1/$+69.2\%$
	+ DCT	**45.4**	52.1	36.9	44.8/$+3.3\%$	**32.1**	**29.5**	**7.1**	**15.2**	**21.0**/$+85.6\%$
Qwen2.5-3B	Direct Inference	10.6	28.8	10.8	16.7	14.9	24.4	2.0	2.4	10.9
	Search-o1	15.1	44.3	13.1	24.2	18.7	17.6	5.8	29.6	17.9
	ReAct	21.1	43.5	28.3	31.0	19.2	19.1	4.8	20.0	15.8
	+ GRPO	44.4	58.0	42.0	48.1/Δ_{base}	39.0	36.3	15.2	36.8	31.8/Δ_{base}
	+ GSPO	43.0	58.8	42.5	48.1/-0.1%	40.2	39.8	17.0	36.8	33.5/$+5.1\%$
	+ Tree-GRPO	46.8	**59.7**	**43.6**	**50.0**/$+3.9\%$	42.4	43.7	17.8	43.2	36.8/$+15.6\%$
	+ DCT	**49.2**	56.7	40.6	48.8/$+1.5\%$	**46.9**	**48.8**	**19.2**	**49.1**	**41.0**/$+28.8\%$
Llama3.2-3B	Direct Inference	16.2	29.6	7.4	17.7	12.6	9.2	2.0	8.0	8.0
	Search-o1	24.2	48.4	8.8	27.1	19.4	17.4	6.0	32.0	18.7
	ReAct	23.9	42.4	21.7	29.3	16.2	10.4	3.5	23.2	13.3
	+ GRPO	45.5	58.2	**42.4**	48.7/Δ_{base}	36.0	26.9	11.8	32.0	26.7/Δ_{base}
	+ GSPO	41.2	57.8	40.8	46.6/-4.3%	28.1	24.5	8.6	32.0	23.3/-12.7%
	+ Tree-GRPO	47.7	**59.9**	42.3	**50.0**/$+2.6\%$	44.6	38.4	17.6	46.4	36.8/$+37.8\%$
	+ DCT	**49.6**	55.7	38.1	47.8/-1.8%	**48.8**	**42.5**	**18.5**	**52.1**	**40.5**/$+51.7\%$
Qwen2.5-7B	Direct Inference	13.4	40.8	14.0	22.7	18.3	25.0	3.1	12.0	14.6
	Search-o1	23.8	47.2	26.2	32.4	22.1	21.8	5.4	32.0	20.3
	ReAct	30.6	56.3	34.6	40.5	27.9	25.3	11.3	28.8	23.3
	+ GRPO	45.8	61.5	44.3	50.5/Δ_{base}	42.5	40.7	19.1	43.2	36.4/Δ_{base}
	+ GSPO	47.0	**64.5**	**46.1**	**52.5**/$+4.0\%$	40.0	38.2	19.2	44.0	35.4/-2.8%
	+ Tree-GRPO	48.1	63.3	45.2	52.2/$+3.3\%$	44.6	42.3	20.2	44.0	37.8/$+3.8\%$
	+ DCT	**49.4**	62.6	44.3	52.1/$+3.1\%$	**47.7**	**45.5**	**21.9**	**47.9**	**40.8**/$+12.0\%$
Qwen2.5-14B	Direct Inference	19.8	53.1	18.4	30.4	21.7	25.3	4.5	16.0	16.9
	Search-o1	34.7	63.5	24.1	40.8	26.8	16.1	9.9	41.6	23.6
	ReAct	36.1	64.2	39.3	46.5	39.1	33.8	15.0	43.2	32.8
	+ GRPO	51.3	67.2	46.7	55.1/Δ_{base}	47.7	42.6	23.2	53.6	41.8/Δ_{base}
	+ GSPO	50.7	67.4	47.1	55.1/$+0.0\%$	50.1	50.2	23.8	52.8	44.2/$+5.9\%$
	+ Tree-GRPO	51.7	**68.1**	**47.3**	**55.7**/$+1.2\%$	50.2	50.5	25.9	54.4	45.3/$+8.3\%$
	+ DCT	**52.5**	66.9	46.5	55.3/$+0.4\%$	**52.6**	**53.2**	**27.1**	**57.7**	**47.7**/$+14.1\%$

6 Experiments

6.1 Experimental Setup

Datasets. We use NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), and PopQA (Mallen et al., 2023) for Single-Hop QA, and HotpotQA (Yang et al., 2018), 2WikiMultiHopQA (Ho et al., 2020), MuSiQue (Trivedi et al., 2022), and Bamboogle (Press et al., 2023) for Multi-Hop QA.

Models. We evaluate Qwen2.5 (Yang et al., 2024) (1.5B, 3B, 7B, 14B) and Llama3.2-3B (Team, 2024) for cross-family evaluation. Quantitative analyses and ablations use Qwen2.5-3B.

Baselines. We compare with Direct Inference, Search-o1 (Li et al., 2025a), ReAct (Yao et al., 2023), and RL baselines GRPO (DeepSeek-AI, 2025), GSPO (Zheng et al., 2025), and Tree-GRPO (Ji et al., 2026), the closest tree-based method.

Experimental settings. We follow the prompts and interaction protocol of Search-R1 (Jin et al., 2025), using a fixed E5 retriever (Wang et al., 2022) over the DPR Wikipedia corpus (Karpukhin et al., 2020). Comparable RL methods share the training data, prompts, rewards, interaction limits, optimization settings, and continuation-rollout budgets. DCT freezes the behavior policy during each tree construction. Further implementation details are provided in Appendix C.1.

Metrics. We report Exact Match (EM) and the average number of tool-call actions (Act.). We further evaluate alternative discovery, failure recovery, causal validity, and inference-time path quality. Detailed definitions of Alternative Coverage, Recovery Rate, CF-EM, PNS, positive action reduction, UTR, and RDR are provided in Appendix C.2.

6.2 Main Results

Single-Hop QA. DCT remains competitive on Single-Hop QA but does not consistently outperform Tree-GRPO (Table 1). On Qwen2.5-3B, it achieves 48.8 average EM versus 50.0 for Tree-GRPO and 48.1 for GRPO; the gap to Tree-GRPO narrows to 0.1 and 0.4 points at 7B and 14B, respectively. DCT stays within 2.7 points of the best method across model blocks. Most single-hop questions need only one effective retrieval before answering, making terminal feedback relatively unambiguous and limiting gains from same-history comparison and backward refinement. Allocating rollouts to alternatives may also disrupt sufficient retrieval strategies. These results agree with Tree-GRPO's observation of smaller process-level gains on shallow trajectories.

Multi-Hop QA. By contrast, DCT achieves the highest average Multi-Hop QA EM across Qwen2.5 scales (1.5B–14B) and Llama3.2-3B. On Qwen2.5-3B, it reaches 41.0, exceeding GRPO and Tree-GRPO by 9.2 and 4.2 points, respectively. Beyond Tree-GRPO's gains over chain-based RL, DCT benefits from same-history action comparisons rather than uniform trajectory-level credit. The contrast with Single-Hop QA suggests that DCT helps most when longer sequences of retrieval and reasoning decisions create greater credit ambiguity. For RL methods, Δ_{rel} denotes the relative change in average EM versus GRPO under the same base model.

6.3 Quantitative Analysis

6.3.1 Training Dynamics

DCT maintains the highest training reward after the initial stage (Figure 3). At update 180, it reaches 39.5 validation EM, exceeding Tree-GRPO (36.8), GSPO (33.5), and GRPO (31.8). Despite non-monotonic validation curves, its strongest endpoint supports transfer to held-out questions. The axis counts policy updates, measuring optimization progress rather than wall-clock efficiency.

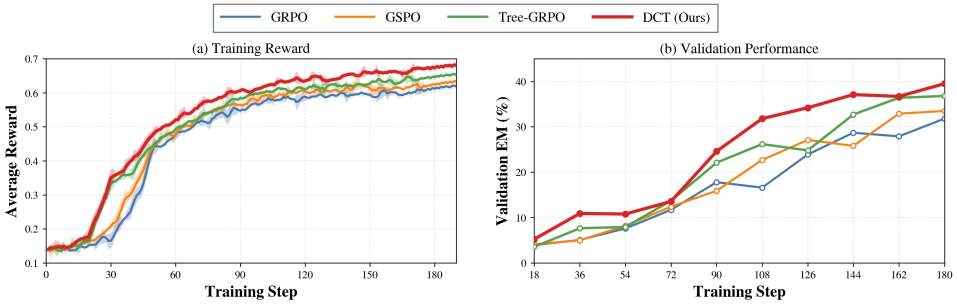

Figure 3: Qwen2.5-3B training reward and validation EM on Multi-Hop QA across policy updates.

6.3.2 Alternative Discovery and Failure Recovery

DCT achieves higher Alternative Coverage and Recovery Rate than both baselines (Figure 4). Coverage rises from 28.3% at step 1 to 43.2% at step 3, versus 29.0% for Tree-GRPO and 31.2% for DCT-Random. Recovery reaches 37.8% at step 1, versus 23.7% and 18.5%, and 20.3% at step 4, versus 12.4% and 9.8%, respectively. Since DCT-Random matches DCT's expansion count, these gains support relevance-guided history selection rather than additional sampling.

6.3.3 Performance by Interaction Demand

We group questions by the shortest successful action count in independent reference trajectories, excluding those without a successful reference and pooling counts ≥ 5. Multi-Path Ratio is the percentage of questions with ≥ 2 distinct successful reference trajectories.

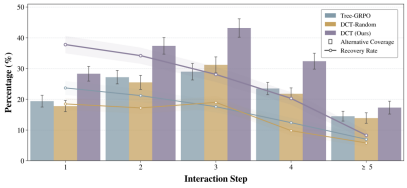

Figure 4: Alternative Coverage (bars) and Recovery Rate (lines) across interaction steps.

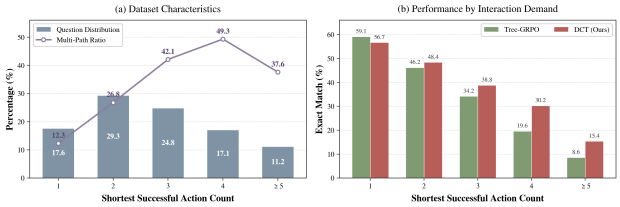

Figure 5: (a) Question distribution and Multi-Path Ratio (b) Tree-GRPO/DCT EM by shortest successful action count.

Table 2: Held-out interventions. CF-EM: replacements preserving correctness (%); PNS: estimated on successful factual trajectories; $\Delta\mathrm{Act}^{+}$: maximum action reduction from successful replacements.

	CF-EM (%)				PNS (%)	$\Delta\mathrm{Act}^{+}$
Intervention Target	Hotpot	2Wiki	MuSiQ	Bamb.		
Random	38.5	44.0	18.5	26.0	68.2	0.58
Tree-GRPO High Credit	33.1	35.0	11.5	21.5	74.7	0.51
DCT Replaceable	73.0	57.5	56.2	40.6	43.2	0.95
DCT Protected	**33.0**	**32.0**	**10.5**	**16.0**	**77.1**	**0.38**

Figure 5(a) shows that 71.2% of questions require two to four actions, with two actions most common (29.3%). Multi-Path Ratio rises from 12.3% for one action to 42.1% and 49.3% for three and four actions, respectively. It remains 37.6% for the ≥ 5 group, which accounts for 11.2% of questions.

Figure 5(b) shows DCT trailing Tree-GRPO on one-action questions (56.7 vs. 59.1 EM), where few intermediate decisions limit comparison and refinement. DCT gains 2.2, 4.6, and 10.6 points for two, three, and four actions, respectively, and 6.8 points for ≥ 5 actions (15.4 vs. 8.6 EM). Greater successful-path diversity aligns with larger gains, supporting same-history comparisons on multi-step tasks and helping explain DCT's Multi-Hop QA advantage.

6.3.4 Causal Validity of Decision Estimates

We independently audit decisions identified during training by replacing actions at fixed histories and regenerating continuations. DCT-Replaceable contains decisions with non-inferior alternatives under the candidate-margin test; DCT-Protected contains those preferred over sampled alternatives.

On held-out rollouts (Table 2), CF-EM is 56.8% for DCT-Replaceable versus 22.9% for DCT-Protected, a 33.9-point gap supporting the candidate-margin distinction. Successful replacements in the former group yield an average maximum reduction of 0.95 actions. Since factual trajectories are correct, PNS complements CF-EM here: DCT-Protected's 77.1% PNS indicates that most replacements cause failure. Tree-GRPO's high-credit group resembles DCT-Protected, suggesting that trajectory credit tracks success without explicitly identifying safe replacements.

6.3.5 Inference-Time Reasoning Quality

All methods generate one trajectory under identical decoding settings; DCT uses its final learned policy without inference-time tree construction or counterfactual expansion. Table 3 shows that ReAct uses the fewest actions but achieves only 15.8 EM and the highest UTR, suggesting premature termination. DCT reaches 41.0 EM with 2.42 actions, versus Tree-GRPO's 36.8 EM with 2.86, and lowers UTR and RDR to 12.8% and 11.8%, respectively. These results indicate reduced redundancy while preserving interactions needed for correct answers.

6.3.6 Ablation Study

Tree construction and reference refinement. Random expansion lowers EM from 40.7 to 37.8 (Table 4), supporting relevance-guided selection. Removing the Margin Gate yields the shortest

Table 3: Multi-Hop QA path quality. Act.: mean tool calls; UTR/RDR: underthinking/redundant-decision rates(%).

Method	EM ↑	Act.	UTR ↓	RDR ↓
ReAct	15.8	2.14	35.4	21.6
GRPO	31.8	2.43	25.7	30.2
GSPO	33.5	2.51	23.8	27.9
Tree-GRPO	36.8	2.86	19.6	20.5
DCT	**41.0**	**2.42**	**12.8**	**11.8**

Table 4: DCT ablations on Qwen2.5-3B: EM (%) and Act. (matched-seed validation split).

Method	Hotpot	2Wiki	MuSiQ	Bamb.	Avg. ↑	Act.
Tree-GRPO	42.4	43.7	17.8	43.2	36.8	2.37
Random Expansion	43.8	44.6	**18.1**	44.7	37.8	1.95
w/o Margin Gate	44.7	45.7	17.2	46.1	38.4	1.58
w/o Length Gate	43.9	45.0	17.3	44.9	37.8	2.11
w/o Ref. Refresh	45.8	47.3	17.6	47.6	39.6	2.03
Full DCT	**46.9**	**48.8**	17.9	**49.1**	**40.7**	**1.78**

paths but only 38.4 EM, suggesting that aggressive replacement may discard necessary decisions. Removing the Length Gate or Reference Refresh lowers accuracy and lengthens paths. Full DCT combines targeted expansion and reference refinement for accurate, concise trajectories.

Process-credit construction. With trees and budget fixed (Table 5), Trajectory-Level Credit performs worst, ignoring local effects. Mean-Including-Focal Credit shrinks contrasts (39.9 EM); Leaf-Count-Weighted Credit is sensitive to uneven expansion; Path-Copy Updates repeat shared-prefix updates. Full DCT improves average EM over the four ablated variants by 0.8–3.1 percentage points and performs best on all benchmarks. Its 40.7 EM supports leave-one-out comparison and anchor-only updates.

Table 5: Process-credit ablations on Qwen2.5-3B with matched search and training settings.

Credit Assignment	Hotpot	2Wiki	MuSiQ	Bamb.	Avg. EM ↑
Trajectory-Level Credit	43.5	45.0	16.6	45.3	37.6
Mean-Including-Focal Credit	46.1	47.8	17.4	48.2	39.9
Leaf-Count-Weighted Credit	45.8	47.4	17.5	47.7	39.6
Path-Copy Updates	44.6	46.4	17.0	46.4	38.6
Full DCT Credit	**46.9**	**48.8**	**17.9**	**49.1**	**40.7**

Tree allocation. Under $ML = 8$, increasing L from one to four raises EM from 33.9 to 40.7, while using a single tree reduces anchor coverage and EM to 39.5 (Table 6). With $M = 2$, EM increases from 37.6 to 40.7 as L grows from two to four, but remains at 40.5 and 40.1 for $L = 6$ and $L = 8$. This diminishing return indicates that additional within-tree trajectories cannot compensate for reduced anchor diversity. Thus, $M = 2, L = 4$ balances independent anchor coverage and within-tree intervention evidence.

Table 6: Tree configurations on Qwen2.5-3B: M anchor trees, $L = B + 1$ terminal trajectories.

Configuration	Hotpot	2Wiki	MuSiQ	Bamb.	Avg. EM ↑
$M = 8,\ L = 1$	40.8	39.6	16.1	39.2	33.9
$M = 4,\ L = 2$	44.8	46.0	17.2	46.4	38.6
$M = 2,\ L = 4$	**46.9**	**48.8**	**17.9**	**49.1**	**40.7**
$M = 1,\ L = 8$	45.6	47.4	17.6	47.5	39.5
$M = 2,\ L = 2$	43.7	44.9	16.7	45.1	37.6
$M = 2,\ L = 3$	45.5	47.1	17.4	47.7	39.4
$M = 2,\ L = 6$	46.8	48.6	17.8	48.8	40.5
$M = 2,\ L = 8$	46.4	48.2	17.7	48.1	40.1

7 Conclusion

DCT separates causal relevance from signed credit through shared-prefix interventions, improving multi-hop accuracy and reasoning efficiency while using only the learned policy at inference.

AI USE STATEMENT

Generative AI tools were used during manuscript preparation for language refinement, structural editing, and discussion of presentation and formulation. All technical claims, mathematical derivations, citations, experimental procedures, and final content are subject to author verification, and the authors take full responsibility for the submitted manuscript.

REFERENCES

Mingyang Chen, Linzhuang Sun, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z. Pan, Wen Zhang, Huajun Chen, Fan Yang, Zenan Zhou, and Weipeng Chen. Research: Learning to reason with search for llms via reinforcement learning. In *Proceedings of Advances in Neural Information Processing Systems*, 2025.

Jie Cheng, Gang Xiong, Ruixi Qiao, Lijun Li, Chao Guo, Junle Wang, Yisheng Lv, and Fei-Yue Wang. Stop summation: Min-form credit assignment is all process reward model needs for reasoning. In *Proceedings of Advances in Neural Information Processing Systems*, 2025.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948, 2025.

Jiazhan Feng, Shijue Huang, Xingwei Qu, Ge Zhang, Yujia Qin, Baoquan Zhong, Chengquan Jiang, Jinxin Chi, and Wanjun Zhong. Retool: Reinforcement learning for strategic tool use in LLMs. In *Proceedings of International Conference on Learning Representations*, 2026.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for LLM agent training. In *Proceedings of Advances in Neural Information Processing Systems*, 2025.

Yiran Guo, Lijie Xu, Ji Liu, Dan Ye, and Shuang Qiu. Segment policy optimization: Effective segment-level credit assignment in RL for large language models. In *Proceedings of Advances in Neural Information Processing Systems*, 2025.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing A multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of International Conference on Computational Linguistics*, pp. 6609–6625, 2020.

Chenghua Huang, Lu Wang, Fangkai Yang, Pu Zhao, Qingwei Lin, Dongmei Zhang, and Saravan Rajmohan. Pretrain value, not reward: Decoupled value policy optimization. In *Proceedings of International Conference on Learning Representations*, 2026.

Yuxiang Ji, Ziyu Ma, Yong Wang, Guanhua Chen, Xiangxiang Chu, and Liaoni Wu. Tree search for LLM agent reinforcement learning. In *Proceedings of International Conference on Learning Representations*, 2026.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan O Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training LLMs to reason and leverage search engines with reinforcement learning. In *Proceedings of Conference on Language Modeling*, 2025.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of Association for Computational Linguistics*, pp. 1601–1611, 2017.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pp. 6769–6781, 2020.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron C. Courville, and Nicolas Le Roux. Vineppo: Refining credit assignment in RL training of llms. In *Proceedings of International Conference on Machine Learning*, 2025.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466, 2019.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. Search-o1: Agentic search-enhanced large reasoning models. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pp. 5420–5438, 2025a.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yongkang Wu, Ji-Rong Wen, Yutao Zhu, and Zhicheng Dou. Webthinker: Empowering large reasoning models with deep research capability. In *Proceedings of Advances in Neural Information Processing Systems*, 2025b.

Jiashun Liu, Johan S. Obando-Ceron, Han Lu, Yancheng He, Weixun Wang, Wenbo Su, Bo Zheng, Pablo Samuel Castro, Aaron C. Courville, and Ling Pan. Asymmetric proximal policy optimization: mini-critics boost LLM reasoning. In *Proceedings of International Conference on Learning Representations*, 2026.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of Association for Computational Linguistics*, pp. 9802–9822, 2023.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of Association for Computational Linguistics: EMNLP*, pp. 5687–5711, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017.

Llama Team. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Musique: Multihop questions via single-hop question composition. *Trans. Assoc. Comput. Linguistics*, 10:539–554, 2022.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *CoRR*, abs/2212.03533, 2022.

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

Zhiheng Xi, Chenyang Liao, Guanyu Li, Zhihao Zhang, Wenxiang Chen, Binghai Wang, Senjie Jin, Yuhao Zhou, Jian Guan, Wei Wu, Tao Ji, Tao Gui, Qi Zhang, and Xuanjing Huang. Agentprm: Process reward models for LLM agents via step-wise promise and progress. In *Proceedings of the ACM World Wide Web Conference*, 2026b.

Yutao Xie, Nathaniel Thomas, Nicklas Hansen, Yang Fu, Li Erran Li, and Xiaolong Wang. TIPS: Turn-level information-potential reward shaping for search-augmented LLMs. In *Proceedings of International Conference on Learning Representations*, 2026.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *CoRR*, abs/2412.15115, 2024.

Matthew Y. R. Yang, Hao Bai, Ian Wu, Gene Yang, Amrith Setlur, and Aviral Kumar. Int: Self-proposed interventions enable credit assignment in LLM reasoning. In *Proceedings of International Conference on Learning Representations*, 2026.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *Proceedings of International Conference on Learning Representations*, 2023.

Haozhen Zhang, Tao Feng, and Jiaxuan You. Router-r1: Teaching llms multi-round routing and aggregation via reinforcement learning. In *Proceedings of Advances in Neural Information Processing Systems*, 2025a.

Wenlin Zhang, Xiangyang Li, Kuicai Dong, Yichao Wang, Pengyue Jia, Xiaopeng Li, Yingyi Zhang, Derong Xu, Zhaocheng Du, Huifeng Guo, Ruiming Tang, and Xiangyu Zhao. Process vs. outcome reward: Which is better for agentic RAG reinforcement learning. In *Proceedings of Advances in Neural Information Processing Systems*, 2025b.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization. *CoRR*, abs/2507.18071, 2025.

A Supplementary Definitions

This appendix specifies the objects and assumptions used in Sections 3–5. We distinguish population intervention targets from the finite evidence recorded by DCT.

A.1 Interaction and Tree Quantities

Definition 3 (Decision and replay context)**.** *At turn t, $H_t = (x, O_{\le t}, A_{<t})$ is the information available to the policy, and A_t is one complete turn-level decision. A replay context Z_t contains the pre-decision environment state and execution metadata needed to resume the interaction at H_t. It may contain information not present in the policy-visible history.*

We suppress t when it is determined by the history. DCT stores the association between a history and its replay context. Histories with different stored contexts are not merged merely because their visible text is identical. For notational simplicity, a tree history h also identifies its stored occurrence when discussing implementation.

Definition 4 (Anchor and intervention tree)**.** *An anchor τ_m^0 is a complete trajectory sampled under π_{old} before adaptive tree expansion. Its interaction histories and actions define the factual path in $\mathcal{T}_m = (\mathcal{H}_m, \mathcal{E}_m)$. Nodes represent histories; outgoing edges represent actions followed by the resulting observations. The tree stores all generated continuations, including failed and unselected ones. A selected successful path is a search reference and does not replace τ_m^0 in the actor-update batch.*

Definition 5 (Action draws and value samples)**.** *For a stored history h, let $\mathcal{D}(h) = (a_1, \ldots, a_{K(h)})$ be the ordered list of policy action draws, with multiplicity retained, and let $\mathcal{A}_{\mathcal{T}}(h) = \{a_j : 1 \le j \le K(h)\}$ be its set of distinct actions. The value ledger for (h, a) contains the terminal returns of fresh continuations that actually execute that history–action occurrence. Its sample count is $N(h, a)$. Different draws of the same action can share this value ledger while remaining separate entries in $\mathcal{D}(h)$.*

Thus, $K(h)$ counts action draws, not distinct children or descendant leaves. Replaying a stored prefix does not create another action draw or value observation for its ancestor edges. These accounting conventions are made explicit in Appendix C.3.

A.2 Same-History Interventions and Replay

Definition 6 (Same-history intervention)**.** *For a feasible action $a \in \mathcal{A}(h)$, the intervention $\mathrm{do}(A_t = a)$ holds the pre-decision history and context fixed, replaces the decision mechanism at turn t, and resumes the environment and policy mechanisms afterward. All subsequent decisions follow the same downstream policy π. This operation changes the current decision without optimizing or fixing the future suffix.*

The distinction between a stored context and a history-conditioned population is important. Write

$$q_t^\pi(h, a; z) := \mathbb{E}[R \mid H_t = h, Z_t = z, \mathrm{do}(A_t = a)], \qquad Q_t^\pi(h, a) = \mathbb{E}_{Z_t \mid H_t = h}[q_t^\pi(h, a; Z_t)]. \tag{A.1}$$

Repeated continuations from one saved context estimate $q_t^\pi(h, a; z)$. They estimate $Q_t^\pi(h, a)$ directly when the relevant context is determined by h, or when the pre-decision contexts are independently sampled from their conditional law. Fresh downstream randomness alone does not resample the pre-decision context. Averaging over correctly distributed anchor contexts recovers the population target under the fixed-design conditions in Appendix B.4.

Definition 7 (Suffix replacement)**.** *Let τ^{ref} be a successful reference and let h_t lie on that path. Replacing its suffix by an observed continuation from h_t preserves the prefix before the decision at t. The original suffix remains in the evidence tree. Replacement therefore changes the search reference but neither deletes evidence nor changes the anchor-update distribution.*

A.3 Credit and Reference Statistics

The population quantities Q_t^π, V_t^π, Δ_t^π, A_t^π, and Γ_t^π are defined in Eqs. 4 and 9. Their finite-tree counterparts have different comparison sets:

- $\widehat{\Gamma}_{\mathcal{T}}(h, a_j)$ compares against distinct sampled actions, retaining the largest absolute contrast.
- $\widehat{A}_{\mathcal{T}}(h, a_j)$ uses a leave-one-out mean over the other *draw indices*. Repeated draws of a_j remain in this baseline when their indices differ from j.
- $\widehat{m}_{\mathcal{T}}(h, a_j)$ compares the reference with the best-valued distinct sampled replacement. It is a signed, candidate-relative test and is not a global necessity certificate.

When no distinct replacement is available, relevance and the replacement margin are unassessed. In particular, an unassessed relevance score is not evidence of causal irrelevance. DCT may use zero as its scheduling default, as defined in Appendix C.3.2, while retaining the unassessed flag. For a factual anchor decision with $K(h) < 2$, the fallback credit is that anchor's original terminal reward.

Definition 8 (Reference length and rollout budget)**.** *The reference length $c(\tau)$ counts interaction turns, including the terminal decision under the same convention for every path. Let $n_{\mathrm{tool}}(\tau)$ count tool invocations separately; these two counts need not coincide. Each anchor receives at most B additional complete continuation rollouts. The realized number of sampled terminal trajectories is $L_m = 1 + B_m$ with $0 \le B_m \le B$, and the nominal per-tree cap is $L = B + 1$.*

Neither c nor n_{tool} augments the task reward. The length criterion selects or refines an observed successful reference only. Counting complete continuations also does not equate their token or wall-clock costs.

A.4 Assumptions

Assumption 9 (Sequential randomization and invariant mechanisms)**.** *The stochastic mechanisms in Eq. 3 are understood with explicit exogenous inputs:*

$$\begin{aligned} S_1 &= f_\rho(x, U_1^S), & O_t &= f_\Omega(S_t, U_t^O), \\ A_t &= f_\pi(H_t, U_t^A; \theta), & S_{t+1} &= f_P(S_t, A_t, U_{t+1}^S). \end{aligned} \tag{A.2}$$

The exogenous innovations are mutually independent conditional on the fixed task. In particular, the policy samples A_t using only H_t and fresh policy randomness. Environment, observation, and reward mechanisms are invariant to the decision intervention and do not depend on θ.

This assumption justifies the identification of same-history interventional values with observational action values on policy support. It does not assert that the history reveals the full environment state.

Assumption 10 (Finite horizon and gradient regularity)**.** *Rewards lie in* $[0, 1]$ *and interactions terminate by a finite* $T_{\max}$*. The policy is differentiable with locally unchanged support, differentiation can be exchanged with integration, and score-function gradients have finite second moments. Histories after termination are padded with an absorbing action whose score is zero.*

Absorbing padding allows the random-horizon sums in the main text to be written as fixed sums through $T_{\max}$. Bounded rewards and finite second moments suffice for the credit-error bound.

Assumption 11 (Faithful replay)**.** *A saved context restores the correct pre-decision state. Every new continuation uses fresh downstream policy and environment randomness, with* π_{old} *held fixed. Population claims average contexts according to* $P(Z_t \mid H_t = h)$*; claims conditional on a stored context refer instead to* $q_t^{\pi_{\text{old}}}(h, a; z)$*.*

The assumption must hold for the environment used for intervention collection; it is not implied by restoring a text prompt alone.

Assumption 12 (Feasible actions and fixed-design estimation)**.** *The feasible intervention set contains the support of the downstream policy. Statements explicitly described as fixed-design guarantees use a preassigned number of independent policy action draws and continuation counts that are not chosen from their terminal returns. Conditional on the action draws and context, each edge-value estimate has the correct conditional mean.*

This last assumption is used only for the fixed-design estimation results. DCT's adaptive scheduler need not satisfy it; its residual estimation error is handled by the general error-transfer bound. Sampling only from π_{old} does not guarantee coverage of feasible interventions outside its support.

B Proofs

We provide the identification and policy-gradient arguments underlying Section 4, prove Proposition 1, and distinguish the population targets from the finite, adaptively collected estimates used by DCT. Throughout, a finite interaction is padded with absorbing states up to $T_{\max}$, with zero scores after termination. We assume the usual differentiation–integration interchange, policy-independent environment mechanisms, and square-integrable score functions. Rewards lie in $[0, 1]$. All claims about unsupported interventions refer to the specified SCM, rather than observational identification.

B.1 Interventional Identification and the Policy Gradient

Same-history identification. Let Z_t denote the complete pre-decision environment context needed for replay, including any latent state not determined by H_t. The sequential SCM samples the action using fresh policy randomness independent of Z_t and future environment randomness conditional on H_t. Consequently,

$$P(dz \mid H_t = h, A_t = a) = P(dz \mid H_t = h) \tag{B.1}$$

for actions in the policy support. Write

$$q_t^\pi(h, a; z) := \mathbb{E}[R \mid H_t = h, Z_t = z, \mathrm{do}(A_t = a)],$$

where later actions follow π. Replacing the action mechanism does not change the pre-decision context distribution. Consistency and Eq. B.1 therefore give

$$\begin{aligned} Q_t^\pi(h, a) &= \int q_t^\pi(h, a; z)\, P(dz \mid H_t = h) \\ &= \mathbb{E}[R \mid H_t = h, A_t = a]. \end{aligned} \tag{B.2}$$

For continuous action spaces, this equality is understood almost everywhere with respect to the policy and appropriate regular conditional distributions. It need not identify actions outside policy support. Latent environment state does not itself invalidate the identity; the relevant condition is that the policy uses H_t and independent action randomness, rather than additional latent information.

Policy-compatible credit. Only policy factors in Eq. 1 depend on θ. Differentiating the trajectory likelihood yields

$$\nabla_\theta J(\theta) = \mathbb{E}\left[R \sum_{t=1}^{T_{\max}} s_t \right].$$

Because s_t is measurable with respect to (H_t, A_t), the tower property and Eq. B.2 replace R by $Q_t^\pi(H_t, A_t)$ inside each expectation. Moreover,

$$\mathbb{E}[s_t \mid H_t = h] = \int \pi_\theta(a \mid h) \nabla_\theta \log \pi_\theta(a \mid h)\, da = \nabla_\theta \int \pi_\theta(a \mid h)\, da = 0.$$

Subtracting the history-dependent baseline $V_t^\pi(H_t)$ establishes Eq. 5. The same argument applies to any integrable history-dependent baseline, treated as a stop-gradient weight within the score-function update.

B.2 Proof of the Credit Misalignment Bound

Proof of Proposition 1. Let $e_t = \widehat{A}_t - A_t^\pi(H_t, A_t)$. Subtracting Eq. 5 from the expected estimated-credit update gives

$$\widehat{g} - \nabla_\theta J = \sum_{t=1}^{T_{\max}} \mathbb{E}[s_t e_t].$$

The triangle inequality, Jensen's inequality for the norm, and Cauchy–Schwarz imply

$$\begin{aligned} \|\widehat{g} - \nabla_\theta J\| &\le \sum_t \|\mathbb{E}[s_t e_t]\| \le \sum_t \mathbb{E}[\|s_t\|\, |e_t|] \\ &\le \sum_t \sqrt{\mathbb{E}\|s_t\|^2}\, \sqrt{\mathbb{E}[e_t^2]}. \end{aligned}$$

The assumed second moments make every term finite. No independence between s_t, $\widehat{A}_t$, and the search outcomes is required. □

What the bound measures. Only credit error correlated with the score can change the expected update. In particular, adding $b_t(H_t)$ to a credit leaves its expected update unchanged even though its pointwise error can increase. More precisely, defining

$$\bar{e}_t(h, a) := \mathbb{E}[\widehat{A}_t \mid H_t = h, A_t = a] - A_t^\pi(h, a)$$

gives the exact identity $\widehat{g} - \nabla J = \sum_t \mathbb{E}[s_t \bar{e}_t(H_t, A_t)]$. The proposition remains valid with e_t replaced by $\bar{e}_t$ and, by conditional Jensen, the latter bound is no larger. Thus the result is an error-control bound, not a claim that every nonzero credit error necessarily changes the gradient.

B.3 Complementary Roles of Relevance and Signed Credit

Suppose the feasible intervention set contains the policy support. For a singleton feasible set we adopt the convention $\Gamma_t^\pi = 0$. From the definitions,

$$A_t^\pi(h, a) = \mathbb{E}_{a' \sim \pi(\cdot|h)}[\Delta_t^\pi(h; a, a')], \qquad |A_t^\pi(h, a)| \le \Gamma_t^\pi(h, a). \tag{B.3}$$

A supremum of nonnegative absolute contrasts is zero exactly when every feasible contrast is zero. This proves the stated invariance interpretation of Γ_t^π and implies $\Gamma_t^\pi = 0 \Rightarrow A_t^\pi = 0$ under the support condition. The reverse implication does not follow: the positive and negative parts of $\Delta_t^\pi(h; a, A')$ can have equal expectations while neither vanishes. In addition, interventions outside policy support can have nonzero effects without entering the policy-weighted average. Therefore the canonical pair (Γ_t^π, A_t^π) satisfies the two population requirements for different reasons: the first by its zero-set definition, and the second by the policy-gradient identity. Neither statement asserts exact relevance faithfulness for a finite sampled tree.

B.4 Analysis of Tree-Based Estimation

Replay targets and latent contexts. Fix h and suppress the turn index. If every continuation independently resamples a context from $P(Z \mid H = h)$ before the intervention, its mean is $Q^\pi(h, a)$. If replay instead restores one saved context z, independent downstream rollouts have mean $q^\pi(h, a; z)$. They are conditionally independent given (h, z), but need not be independent conditional on h alone. In particular, increasing the number of rollouts from a single snapshot removes downstream Monte

Carlo noise; it does not remove variation across latent contexts. Averaging over unselected anchor contexts recovers Q^π by Eq. B.2.

For the following calculations, let $\mu(a)$ denote the mean under the reset protocol: either $Q^{\pi_{\text{old}}}(h, a)$ for independent conditional-context replay, or $q^{\pi_{\text{old}}}(h, a; z)$ conditional on a saved snapshot. Let $v_\mu = \mathbb{E}_{a\sim\pi_{\text{old}}(\cdot|h)}[\mu(a)]$ and $A_\mu(a) = \mu(a) - v_\mu$. The analysis applies only to genuine continuations following π_{old} after the intervened action. A suffix regenerated below an ancestor cannot be counted as another unmodified continuation of that ancestor merely because its prefix was replayed.

A fixed-design leave-one-out benchmark. Take a fixed $K \geq 2$ and draw $a_1, \ldots, a_K$ independently from $\pi_{\text{old}}(\cdot \mid h)$, retaining repeated draws. Conditional on these draws, suppose each evaluated action has an unbiased value estimate, with sample counts chosen independently of continuation returns. Independence between estimates for different actions is not needed. For any predetermined index j,

$$\begin{aligned}\mathbb{E}[\widehat{A}_{\mathcal{T}}(h, a_j) \mid a_j = a] &= \mu(a) - \mathbb{E}\left[\frac{1}{K-1}\sum_{\ell\neq j}\mu(a_\ell)\right] \\ &= \mu(a) - v_\mu = A_\mu(a).\end{aligned} \tag{B.4}$$

Here the expectation is conditional on the fixed history and, when applicable, the replay snapshot. A repeated action can reuse the same value estimate in several terms; it retains its multiplicity in the baseline. Deduplicating the baseline would generally change its target. For saved-context replay, integrating Eq. B.4 over $P(Z \mid H = h, A_j = a) = P(Z \mid H = h)$ recovers $A^{\pi_{\text{old}}}(h, a)$. This is a fixed-design benchmark, not an unbiasedness assertion for the final adaptively stopped DCT estimate.

Finite-sample value errors. Conditional on the action draws, let $\mathcal{S}$ be their d distinct actions and let action a have $n_a \geq 1$ conditionally independent returns of mean $\mu(a)$. For $0 < \delta_Q < 1$, Hoeffding's inequality and a union bound give, with probability at least $1 - \delta_Q$,

$$|\widehat{Q}_{\mathcal{T}}(h, a) - \mu(a)| \leq r_a, \qquad r_a := \sqrt{\frac{\log(2d/\delta_Q)}{2n_a}}, \qquad a \in \mathcal{S}. \tag{B.5}$$

The union bound does not require independence across different edge estimates. Write $r = \max_{a\in\mathcal{S}} r_a$.

Propagation to relevance and replacement margins. When a distinct alternative is available, define the exact sample-restricted quantities

$$\Gamma_{\mu,\mathcal{S}}(a_j) = \max_{a\in\mathcal{S}\setminus\{a_j\}} |\mu(a_j) - \mu(a)|, \qquad m_{\mu,\mathcal{S}}(a_j) = \mu(a_j) - \max_{a\in\mathcal{S}\setminus\{a_j\}} \mu(a).$$

On the event in Eq. B.5, the absolute-value and maximum maps are nonexpansive, so

$$|\widehat{\Gamma}_{\mathcal{T}} - \Gamma_{\mu,\mathcal{S}}| \leq 2r, \qquad |\widehat{m}_{\mathcal{T}} - m_{\mu,\mathcal{S}}| \leq 2r. \tag{B.6}$$

For example, $\widehat{m}_{\mathcal{T}} > 2r$ certifies a positive sample-restricted population margin on this event; a nonpositive empirical margin alone is only an estimated comparison. DCT uses the latter as a search rule, together with an observed shorter successful suffix. Neither the empirical sign nor the successful suffix proves global causal necessity.

Even when $\mu = Q^{\pi_{\text{old}}}$, finite sampling introduces the additional nonnegative coverage gap

$$\eta_{\mathcal{S}} := \Gamma^{\pi_{\text{old}}}(h, a_j) - \Gamma_{Q,\mathcal{S}}(a_j).$$

Thus $|\widehat{\Gamma}_{\mathcal{T}} - \Gamma^{\pi_{\text{old}}}| \leq 2r + \eta_{\mathcal{S}}$. The noisy estimated maximum need not be a lower bound on population relevance. Policy sampling can close the coverage gap only if relevant alternatives receive policy mass. More precisely, conditional on $a_j = a$, let p_ζ be the policy mass of distinct alternatives whose absolute contrast is within ζ of the supremum. For fixed independent action draws,

$$\Pr(\text{no distinct alternative or } \eta_{\mathcal{S}} > \zeta \mid a_j = a) \leq (1 - p_\zeta)^{K-1}. \tag{B.7}$$

If no distinct alternative is observed, relevance remains unassessed; Eq. B.7 treats that event as a failure to obtain the required coverage. Feasible but unsupported alternatives cannot be recovered by increasing policy-only sampling.

Propagation to signed credit. For fixed K and a predetermined j, the true values of the other action draws lie in $[0,1]$ and have mean v_μ. With additional probability failure at most δ_A,

$$\left| \frac{1}{K-1} \sum_{\ell \neq j} \mu(a_\ell) - v_\mu \right| \leq b_K, \qquad b_K := \sqrt{\frac{\log(2/\delta_A)}{2(K-1)}}.$$

Combining this bound with Eq. B.5 gives

$$|\widehat{A}_{\mathcal{T}}(h, a_j) - A_\mu(a_j)| \leq r_{a_j} + \frac{1}{K-1} \sum_{\ell \neq j} r_{a_\ell} + b_K \leq 2r + b_K \tag{B.8}$$

with probability at least $1-\delta_Q-\delta_A$. The two terms reflect continuation estimation and policy-baseline sampling, respectively. For a saved snapshot, an additional $|A_\mu(a_j) - A^{\pi_{\text{old}}}(h, a_j)|$ is needed for a pointwise bound relative to the history-level target. Similarly, setting $d_z = \max_{a \in \mathcal{S}} |q^{\pi_{\text{old}}}(h, a; z) - Q^{\pi_{\text{old}}}(h, a)|$ adds at most $2d_z$ to the sample-restricted relevance and margin bounds. These context terms cannot be removed by resampling only downstream noise.

Adaptive collection and stopping. DCT uses previous outcomes to choose histories and stopping times; new action draws always follow π_{old}. Fresh random continuations ensure the appropriate conditional sampling law, but a sample mean evaluated at an outcome-dependent stopping time need not be unbiased. Likewise, Eq. B.4 cannot simply be conditioned on a random observed K or on a selected successful reference path. Ordinary fixed-sample confidence bounds cannot be substituted at arbitrary stopping times without an additional argument.

A finite-budget uniform bound is available when each edge has a fixed conditional mean under its reset protocol and each newly counted return is a fresh bounded sample with that mean, conditional on its prior sample stream. If at most $E_{\max}$ edge streams are created and each contributes at most $n_{\max}$ samples, the bounded-increment Hoeffding inequality applied to the potential sample streams and a union bound over streams and sample sizes give the simultaneous radius

$$r^{\text{unif}}(n) = \sqrt{\frac{\log(2E_{\max} n_{\max}/\delta_Q)}{2n}}. \tag{B.9}$$

It holds at the adaptively chosen counts. Stream creation must precede its counted outcomes, and the reset protocol must not change its mean. For factual action index $j = 1$, an analogous union bound over at most $H_{\max}$ histories and $K_{\max} - 1$ subsequent policy draws gives

$$b^{\text{unif}}(K) = \sqrt{\frac{\log(2H_{\max}(K_{\max}-1)/\delta_A)}{2(K-1)}}.$$

Under these sampling conditions the deterministic error propagation in Eqs. B.6 and B.8 remains valid at adaptive counts. This provides confidence control, not exact finite-sample unbiasedness, and does not resolve unsupported alternatives or latent-context error.

Terminal-reward fallback. Using the original anchor reward as the weight at every turn gives Eq. 2. It is therefore a valid Monte Carlo policy-gradient weight, although $\mathbb{E}[R \mid H_t, A_t] = Q_t^\pi$, rather than A_t^π. An adaptively chosen mixture with tree credit requires separate care. For the indicator W_t that tree credit is used, its deviation from the terminal-reward update is exactly

$$\mathbb{E}\left[\sum_t s_t W_t (\widehat{A}_t^{\text{tree}} - R) \right]. \tag{B.10}$$

This need not vanish when W_t depends on actions or observed returns. The validity of the unselected terminal-reward estimator therefore does not, by itself, prove unbiasedness of the adaptive fallback rule.

Transfer to the DCT update. All credit construction, adaptive selection, and fallback effects can be incorporated in the residual error. Since only behavior-policy anchor visits enter Eq. 16, let

$$G_t^2 = \mathbb{E}\|s_t^{\text{old}}\|^2, \qquad \epsilon_t^2 = \mathbb{E}\left[(\widehat{A}_{\mathcal{T}}(H_t, A_t) - A_t^{\pi_{\text{old}}}(H_t, A_t))^2 \right],$$

where expectations include anchor and tree-construction randomness. Proposition 1 immediately yields

$$\|\mathbb{E}[\widehat{g}_{\mathrm{DCT}}] - \nabla_\theta J(\theta_{\mathrm{old}})\| \le \sum_{t=1}^{T_{\max}} G_t \epsilon_t. \tag{B.11}$$

Independent trees are not required for this expectation identity; retaining the anchor visitation law is essential. Where a valid population-target error bound $|\widehat{A} - A| \le u_t$ holds except with probability δ_t, bounded rewards imply $\epsilon_t \le \sqrt{u_t^2 + 4\delta_t}$, since both credits lie in $[-1, 1]$ (the fallback lies in $[0, 1]$). This connects valid finite-sample guarantees to expected-update control without assuming adaptive unbiasedness. Finally, this analysis concerns the update at θ_{old}; using the same weights in a clipped PPO surrogate does not imply equality to the current-policy gradient after multiple optimization steps.

C Implementation Details and Additional Results

This appendix specifies the implementation of Section 5 and the evaluation quantities used in Section 6. Algorithm 1 makes the intervention selection, reference updates, and anchor-only policy update explicit.

C.1 Implementation Details

We evaluate NQ, TriviaQA, and PopQA for Single-Hop QA, and HotpotQA, 2WikiMultiHopQA, MuSiQue, and Bamboogle for Multi-Hop QA. We use Qwen2.5 models at 1.5B, 3B, 7B, and 14B scales and Llama3.2-3B. Unless otherwise specified, quantitative analyses and ablations use Qwen2.5-3B.

Our implementation follows the Search-R1 prompt format and interaction protocol. All interactive methods use the same ReAct-style interface and an E5-based local retriever over the DPR Wikipedia corpus. Comparable RL methods share the training data, prompts, rewards, interaction limits, optimization settings, decoding configuration, and continuation-rollout budgets.

DCT samples M anchor trajectories and at most B additional continuations for each anchor, giving $L = B + 1$ terminal trajectories per tree. The behavior policy is frozen during tree construction. At inference time, DCT uses only the learned policy and does not perform tree construction, counterfactual expansion, or backward refinement.

C.2 Evaluation Metrics and Auditing Protocol

We report normalized Exact Match (EM) and the average number of tool-call actions (Act.). Alternative Coverage is the proportion of audited histories containing at least two distinct actions. Recovery Rate is the proportion of failed decisions for which a successful same-history alternative is found.

For causal evaluation, we preserve the history preceding the evaluated decision and generate independent held-out replacement rollouts. CF-EM is the percentage of replacements that preserve the correct answer, while PNS is the percentage that change a factually correct trajectory into an incorrect one. Under this evaluation protocol, PNS equals one minus CF-EM. Positive Action Reduction measures the largest action reduction obtained by a successful, non-inferior replacement.

Underthinking Rate (UTR) is the proportion of failed trajectories containing a decision with a successful same-history alternative. Redundant-Decision Rate (RDR) is the proportion of audited decisions on successful trajectories that admit a shorter, successful, and non-inferior replacement. All auditing rollouts are independent of those used for training and decision selection.

C.3 Detailed Training Procedure

This section specifies the replay bookkeeping, intervention scheduler, and policy-update interface of DCT. Each behavior-policy anchor receives at most B additional continuation rollouts. The behavior policy π_{old} remains frozen throughout construction of all trees used for one update. The

forward allocation $B_{\text{fwd}} \in \{0, \dots, B\}$ and exploration coefficient $\varepsilon \in (0, 1)$ are fixed configuration parameters.

C.3.1 Replay and Evidence Bookkeeping

For each anchor τ_m^0, DCT stores an evidence tree $\mathcal{T}_m$, replayable pre-decision contexts, and the original terminal return R_m^0. At a history h, it maintains the ordered list of policy-sampled action draws $\mathcal{D}(h) = (a_1, \dots, a_{K(h)})$ and, for each evaluated action, a return sum $S(h, a)$ and continuation count $N(h, a)$. Repeated draws of the same action remain separate entries of $\mathcal{D}(h)$, but use the same action-value estimate $S(h, a)/N(h, a)$. The index of each factual anchor draw is retained for leave-one-out credit assignment.

Expansion operation. An expansion restores the pre-decision context at a selected history h, samples $a' \sim \pi_{\text{old}}(\cdot \mid h)$, appends that draw to $\mathcal{D}(h)$, and executes a complete suffix with fresh downstream randomness under π_{old}. This increases $K(h)$ even when a' duplicates an earlier action. An existing action is re-evaluated when it is drawn again, rather than by separately forcing its selection. The resulting suffix is stored and all affected value and relevance statistics are refreshed before the next intervention is selected.

Restoring one fixed latent state repeatedly estimates values conditional on that state; identification with the history-conditional quantity requires the context-sampling conditions in Appendix A.4. Independent downstream randomness alone does not supply independent samples of the latent pre-decision context.

Return accounting. For a newly executed suffix σ with terminal return r, let $\mathrm{Exec}(\sigma)$ contain its freshly executed history–action occurrences, including the intervention decision. For each $(u, b) \in \mathrm{Exec}(\sigma)$, update

$$S(u, b) \leftarrow S(u, b) + r, \qquad N(u, b) \leftarrow N(u, b) + 1. \tag{C.1}$$

Each freshly sampled downstream decision also appends one draw to $\mathcal{D}_u$. A restored prefix supplies no new ancestor observations: its action-draw lists, counts, and return sums are unchanged. Initialization applies this accounting once to every decision in the anchor. Thus a terminal return is counted once per executed occurrence, and a counterfactual suffix is never backed up through unexecuted ancestor edges as another old-policy continuation.

Failed, replaced, and unselected branches remain available for estimation. Exact action duplicates are merged only for value sums and counts; action-draw multiplicities are never removed from the baseline.

C.3.2 Relevance-Guided Construction

For an eligible history h, let a_h be the first recorded action during forward discovery and the current reference action during backward testing. Define the scheduling score

$$g(h) = \begin{cases} \widehat{\Gamma}_{\mathcal{T}}(h, a_h), & \text{if a distinct alternative has been evaluated,} \\ 0, & \text{otherwise.} \end{cases} \tag{C.2}$$

The zero in the second case is only a scheduling convention, not a conclusion of causal irrelevance. Given a nonempty candidate set $\mathcal{C}$, sample

$$p(h \mid \mathcal{C}) = \begin{cases} \dfrac{\varepsilon}{|\mathcal{C}|} + (1 - \varepsilon)\dfrac{g(h)}{\sum_{u \in \mathcal{C}} g(u)}, & \displaystyle\sum_{u \in \mathcal{C}} g(u) > 0, \\ \dfrac{1}{|\mathcal{C}|}, & \text{otherwise.} \end{cases} \tag{C.3}$$

Every eligible history has positive selection probability. With binary rewards, the scheduler is uniform while all observed returns are zero; relevance-guided allocation becomes active only after sampled outcomes produce value contrasts. The criterion prioritizes detected outcome sensitivity and is not a measure of statistical uncertainty.

Algorithm 1 Dynamic Causal Tree for one policy update

Require: Behavior policy π_{old}; task sampler; M, B, $B_{\text{fwd}} \le B$, ε; replay interface
Ensure: Evidence trees $\{\mathcal{T}_m\}$ and factual-anchor credits $\{\widehat{A}_{m,t}\}$

Freeze π_{old}; sample anchors $\{\tau_m^0\}_{m=1}^M$
for $m = 1, \dots, M$ **do**
  $\mathcal{T}_m \leftarrow \mathsf{Initialize}(\tau_m^0)$; $b \leftarrow B$; $\tau^{\text{ref}} \leftarrow \tau_m^0$ if $R_m^0 = 1$, else $\varnothing$
  if $\tau^{\text{ref}} = \varnothing$ **then**
    for $s = 1, \dots, B_{\text{fwd}}$ **do**
      $\mathcal{C} \leftarrow \mathsf{Eligible}(\mathcal{T}_m)$
      if $\mathcal{C} = \varnothing$ **then**
        break
      end if
      $h \sim p(\cdot \mid \mathcal{C})$ with first-recorded a_h
      $\mathcal{T}_m \leftarrow \mathsf{Expand}(\mathcal{T}_m, h)$; $b \leftarrow b - 1$
    end for
    $\tau^{\text{ref}} \leftarrow \mathsf{ShortestSuccess}(\mathcal{T}_m)$, or $\varnothing$ if none exists
  end if
  while $b > 0$ and $\tau^{\text{ref}} \neq \varnothing$ **do**
    $\mathcal{C} \leftarrow \mathsf{Eligible}(\tau^{\text{ref}})$
    if $\mathcal{C} = \varnothing$ **then**
      break
    end if
    $h \sim p(\cdot \mid \mathcal{C})$ with current-reference a_h
    $\mathcal{T}_m \leftarrow \mathsf{Expand}(\mathcal{T}_m, h)$; $b \leftarrow b - 1$
    Replace the reference suffix if the margin and shorter-success conditions hold; retain all branches
  end while
  Consolidate $\widehat{Q}, \widehat{\Gamma}$ in $\mathcal{T}_m$; assign $\widehat{A}_{m,t}$ to factual anchor draws by Eq. 13, using the fallback where needed
end for
Return $\{\mathcal{T}_m\}$ and $\{\widehat{A}_{m,t}\}$; update the policy with Eq. 16

Forward discovery. A failed anchor enters Stage I with $\mathcal{C}$ equal to the current tree's replayable nonterminal histories. DCT performs up to B_{fwd} expansions using Eq. C.3, adding newly visited eligible histories to $\mathcal{C}$. At the end of that allocation, it selects a shortest observed successful path as in Eq. 14; ties are resolved by a fixed rule. A successful anchor instead enters Stage II directly with its original path as the reference. If Stage I finds no success, Stage II is skipped.

Backward testing. Stage II restricts $\mathcal{C}$ to replayable nonterminal histories of the current successful reference. It uses the same sampling rule, with scores recomputed for the reference decisions. Here "backward" means revisiting decisions on an already completed successful path; histories are not visited in a mandatory reverse chronological order. After each expansion, define the set of best-valued distinct alternatives

$$\mathcal{A}^\star(h, a_h) = \arg \max_{a \in \mathcal{A}_{\mathcal{T}}(h) \setminus \{a_h\}} \widehat{Q}_{\mathcal{T}}(h, a). \tag{C.4}$$

When this set is nonempty, replace the reference suffix only if $\widehat{m}_{\mathcal{T}}(h, a_h) \le 0$ and an action in $\mathcal{A}^\star(h, a_h)$ has an observed successful suffix shorter than the reference suffix. A fixed tie-breaking rule chooses among eligible suffixes. Replacement refreshes the reference histories and actions used for scheduling, but removes no stored evidence. The comparison concerns sampled replacements, not global causal necessity.

Stopping and budget reporting. Construction stops when its assigned budget is exhausted or its candidate set becomes empty. Skipping Stage II leaves its unused allocation unspent. If B_m^{used} continuations are executed, tree m contains $L_m = 1 + B_m^{\text{used}} \le 1 + B$ sampled terminal trajectories, including its anchor. Consequently $L = B + 1$ is a per-tree cap, and is the realized count only when the full allocation is consumed. Budget-matched experiments must report actual continuation counts

or specify the policy for reallocating unused budgets. A continuation budget does not directly match generated-token cost or elapsed time.

C.3.3 From Tree Evidence to Policy Updates

For the factual draw index j at an anchor history h, all other entries of $\mathcal{D}(h)$ contribute to the leave-one-out baseline with their sampling multiplicities, including another draw whose action equals a_j. The distinct-action restriction applies to relevance and replacement testing, not to the baseline. If $K(h) < 2$, the fallback is the original anchor's terminal return, not the return of a later selected reference.

The score-function update in Eq. 16 uses only original anchor visits. In a PPO implementation, a turn-level clipped surrogate is

$$\begin{aligned} \mathcal{L}_{\mathrm{clip}}(\theta) &= \frac{1}{M}\sum_{m=1}^{M}\sum_{t=1}^{T_m}\min\big\{r_{m,t}(\theta)\bar{A}_{m,t},\ \mathrm{clip}\big(r_{m,t}(\theta),1-\epsilon_{\mathrm{clip}},1+\epsilon_{\mathrm{clip}}\big)\bar{A}_{m,t}\big\}, \\ r_{m,t}(\theta) &= \frac{\pi_\theta(A_{m,t}\mid H_{m,t})}{\pi_{\mathrm{old}}(A_{m,t}\mid H_{m,t})}, \qquad \bar{A}_{m,t} = \mathrm{sg}[\widehat{A}_{m,t}]. \end{aligned} \tag{C.5}$$

Only anchor-generated action tokens contribute actor samples; environment observations and counterfactual continuations do not. For an autoregressively generated turn, $\log \pi_\theta(A_t \mid H_t)$ sums the conditional log-probabilities of its generated tokens. Token-level clipping is a separate implementation choice and should be reported explicitly rather than identified with Eq. C.5. No reference-path length penalty is inserted into the task reward.

Adaptive allocation and stopping can affect finite-sample estimates. Correct bookkeeping prevents double-counting but does not by itself make final credit unbiased. The error-transfer analysis and conditions for estimator guarantees are given in Appendix B.4.

D Case Study

Example 1 from 2WikiMultiHopQA: DCT protects an empirically necessary decision

Question: Which country is Edmund I's father from?
Labeled Answer: Wessex; Kingdom of Wessex; Kingdom of the West Saxons
Evaluated Decision: "who was the father of Edmund I of Wessex"

Observed DCT trajectory

<think> Edmund I was a king of Wessex. I need to identify his father and determine where he was from. </think>

<search> who was the father of Edmund I of Wessex </search>

<result> The returned documents identify Edmund I as the son of Edward the Elder and connect Edward's royal lineage to Wessex. *[Remaining supporting passages omitted for space.]* </result>

<answer> Wessex </answer>

Training-time decision: The evaluated search is favored over its sampled same-history candidates by the candidate-margin comparison. No shorter candidate satisfies both the Margin Gate and the Length Gate, so DCT preserves the original reference.

Held-out counterfactual audit

Shortest successful replacement: <answer> Wessex </answer>

Representative failed replacement: <answer> England </answer>

Intervention statistics: Only one of eight held-out replacements remains correct, yielding CF-EM = 12.5%, PNS = 87.5%, and $\Delta\text{Act}^{+} = 1$.

Finding. The held-out audit supports DCT's decision to protect the search. Although one shorter successful alternative is observed, most replacements destroy the correct answer, so replacing the decision would be unreliable.

Table 7: A protected decision with high empirical necessity. Blue, red, green, and purple boxes denote reasoning, search actions, returned evidence, and answers, respectively.

Example 2 from 2WikiMultiHopQA: trajectory-level feedback obscures a redundant decision

Question: Which film was released more recently, *The Face of a Genius* or *Cowboy Bebop: The Movie*?
Labeled Answer: Cowboy Bebop: Knockin' on Heaven's Door; Cowboy Bebop: The Movie; Cowboy Bebop: Tengoku no Tobira
Evaluated Decision: "and"

Observed Tree-GRPO trajectory

<think> I should compare the release years of the two films. </think>

<search> The Face of a Genius release year; Cowboy Bebop: The Movie release year </search>

<result> The returned evidence identifies the release periods of both films and supports that *Cowboy Bebop: The Movie* was released more recently. *[Supporting passages omitted for space.]* </result>

<search> and </search>

<result> The query returns unrelated pages about "and–or trees" and other irrelevant topics. </result>

<answer> Cowboy Bebop: The Movie </answer>

Held-out counterfactual audit

Shortest successful replacement: <answer> Cowboy Bebop: The Movie </answer>

Intervention statistics: All eight held-out replacements preserve the correct answer, yielding CF-EM = 100.0%, PNS = 0.0%, and $\Delta\text{Act}^{+} = 1$.

Finding. The redundant search lies on a successful trajectory and is therefore not distinguished by uniform trajectory-level feedback. Same-history interventions show that the answer is preserved after replacing the search, revealing that the decision contributes no measurable necessity.

Table 8: A redundant decision obscured by trajectory-level feedback. Blue, red, green, and purple boxes denote reasoning, search actions, returned evidence, and answers, respectively.

Example 3 from 2WikiMultiHopQA: a mixed but protected decision

Question: Which film has the younger director, *Dallas 362* or *Revenge of the Barbarians*?
Labeled Answer: Dallas 362
Evaluated Decision: "director age Dallas 362"

Observed DCT trajectory

<think> I need to identify the directors of both films and compare their ages. </think>

<search> director age Dallas 362 </search>

<result> The retrieved evidence identifies Scott Caan as the director of *Dallas 362* and provides the information required for comparison with the director of *Revenge of the Barbarians*. *[Remaining supporting passages omitted for space.]* </result>

<answer> Dallas 362 </answer>

Training-time decision: The sampled alternatives do not provide a sufficiently strong non-inferior replacement under the candidate-margin comparison. The original search is therefore preserved in the reference trajectory.

Held-out counterfactual audit

Shortest successful replacement: <answer> Dallas 362 </answer>

Representative failed replacement: <answer> Revenge of the Barbarians </answer>

Intervention statistics: Two of eight held-out replacements remain correct, yielding CF-EM = 25.0%, PNS = 75.0%, and $\Delta\text{Act}^{+} = 1$.

Finding. Most counterfactual replacements fail, supporting the decision to preserve the search. The mixed outcomes also illustrate that empirical necessity is estimated from finite counterfactual samples rather than treated as an absolute causal label.

Table 9: A protected decision with mixed held-out counterfactual outcomes. Blue, red, green, and purple boxes denote reasoning, search actions, returned evidence, and answers, respectively.

Example 4 from 2WikiMultiHopQA: DCT replaces a high-credit but replaceable search

Question: Which film came out earlier, *Au Revoir Les Enfants* or *Physical Jerks*?
Labeled Answer: Au revoir les enfants; Au Revoir Les Enfants; Goodbye children
Evaluated Decision: "Au Revoir Les Enfants release date"

Factual trajectory containing the evaluated decision

<think> To determine which film appeared earlier, I need to compare their release dates. </think>

<search> Au Revoir Les Enfants release date </search>

<result> The returned evidence identifies *Au Revoir Les Enfants* as a 1987 film and provides the information required for comparison with *Physical Jerks*. *[Remaining supporting passages omitted for space.]* </result>

<answer> Au Revoir Les Enfants </answer>

Trajectory-level attribution: The search is selected by the Tree-GRPO-style high-credit criterion because it lies on a successful trajectory. This attribution, however, does not determine whether the search can be replaced by a shorter successful continuation.

Candidate comparison and reference refresh

Starting from the same preceding history, DCT samples alternative actions and independently completes their continuations. One candidate directly produces the correct answer:

<answer> Au Revoir Les Enfants </answer>

Gate decision: The candidate preserves the terminal outcome and uses one fewer action. It therefore satisfies both the Margin Gate and the Length Gate. DCT classifies the original search as replaceable and refreshes the reference with the complete candidate branch.

Original reference:

<search> Au Revoir Les Enfants release date </search> → <answer> Au Revoir Les Enfants </answer>

Refreshed reference:

<answer> Au Revoir Les Enfants </answer>

DCT replaces the complete continuation rather than deleting the search and reconnecting its original suffix. The refreshed reference therefore remains logically self-contained.

Held-out counterfactual audit

Seven of eight independent replacements preserve the exact answer, and the shortest successful replacement removes one action. The audit therefore yields CF-EM = 87.5%, PNS = 12.5%, and $\Delta\mathrm{Act}^{+} = 1$.

Finding. Trajectory-level attribution assigns high credit to the search because it lies on a successful trajectory. DCT instead compares complete continuations under the same history and discovers a non-inferior, shorter branch. The independent audit confirms that the replaced search has low empirical necessity, while the reference refresh produces a shorter successful trajectory without introducing a discontinuity.

Table 10: DCT replaces a high-credit search after same-history candidate comparison identifies a shorter non-inferior continuation. Held-out interventions independently confirm its low empirical necessity. Blue, red, green, and purple boxes denote reasoning, search actions, returned evidence, and answers, respectively.