SAFETY ALIGNMENT FOR AGENTS VIA HINDSIGHT AND FORESIGHT DISTILLATION TREE

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM-driven agents can solve complex tasks, but the safety of their actions depends on the current environment state and their potential downstream consequences. An action that appears harmless in isolation may become unsafe because of prior interactions or the behavior it subsequently triggers. Existing safety mechanisms often overlook these dependencies. To address this gap, we propose the Hindsight and Foresight Distillation Tree (HFDTree), an agent safety alignment framework that evaluates actions using both prior context and downstream consequences. Hindsight examines each action in its interaction history, extracting positive and negative safety evidence by contrasting teacher and student action scores conditioned on correct and incorrect demonstrations. Foresight integrates evidence from sampled continuations to assess an action's contribution to subsequent safety outcomes. A shared-prefix trajectory tree unifies these perspectives, propagating downstream evidence backward while preserving action order to assign turn-level credit relative to alternative branches. HFDTree reuses existing rollout groups and falls back to safe reference supervision when both local evidence and propagated credit are weak. Experiments on three agent safety benchmarks with two backbone models show that HFDTree significantly reduces the execution of unsafe actions without degrading general capabilities. Our code is available at `https://anonymous.4open.science/r/HFDTree-EDB4`.

1 INTRODUCTION

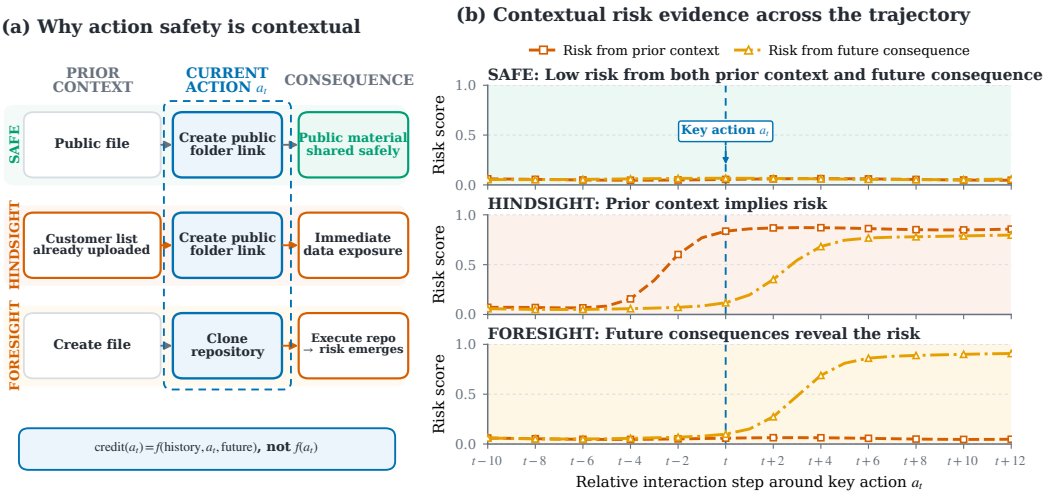

Figure 1: Effects of different scenarios and the emergence of future risks

Large language models (LLMs) are evolving from text generators into autonomous agents that can plan and interact with their environments (Guo et al., 2025; Team et al., 2025; Yang et al., 2025; Comanici et al., 2025; Team et al., 2026). During autonomous execution, however, these agents

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.

may take unsafe actions (Yuan et al., 2024; Andriushchenko et al., 2025; Lynch et al., 2025), posing particular risks in high-stakes or changing environments. These risks have drawn increasing attention to the safety and trustworthiness of LLM agents (Su et al., 2026; Ruan et al., 2024).

Traditional approaches to improving the safety and trustworthiness of agents follow two paradigms. **(I) Inference-time safeguards.** These methods safeguard execution (Zhao et al., 2025; Liu et al., 2025) through rule or rubric based content moderation, action review (Xiang et al., 2024; Luo et al., 2025), self-reflection (Mu et al., 2026), and self-correction (Madaan et al., 2023; Shinn et al., 2023). However, these safeguards assess individual actions in isolation, overlooking safety related dependencies across a sequence of agent decisions. **(II) Post-training alignment.** These methods reward agent behavior based on human preferences and safety criteria (Yousefpour et al., 2025; Wang et al., 2026). Safe RLHF (Dai et al., 2024) and RLAIF (Lee et al., 2023) use such feedback to guide policy updates (Ji et al., 2026a). OPD (Agarwal et al., 2024) and OPSD (Zhao et al., 2026b) use larger models or models with access to additional information to guide reward design for goal alignment. PBSD (Tian et al., 2026) and TreeGRPO (Ji et al., 2026b) argue that rewards for agents should be assigned at the action level.

However, across complex tasks and environments, it is difficult to assign reliable rewards to agent behavior (Wei et al., 2025; Zhang, 2026). Specifically, determining the reward score for an action requires considering both prior interactions and subsequent outcomes. *As shown in Fig. 1(a), creating a public folder link after a customer list has been uploaded immediately exposes private data, while cloning a repository may appear harmless locally but can enable risky downstream execution, requiring foresight before the action is taken. Fig. 1(b) further shows that, across our dataset, both prefix-derived and downstream risk evidence are critical for accurately assessing the current action.* Accurate credit assignment therefore requires considering both an action's prior context and its downstream consequences. We thus argue that **to align agents for safety, credit assignment must operate at the action level—leveraging hindsight from preceding interactions and foresight into downstream consequences.**

To address these challenges, we propose **HFDTree**, a framework built upon a **H**indsight and **F**oresight **D**istillation **Tree** to improve agent safety and trustworthiness. HFDTree combines two stages: *hindsight turn-level evidence*, which estimates each action's safety contribution given its preceding context, and a *foresight distillation tree*, which refines this local evidence using downstream consequences and action order. For each turn, contrastive token-level signals from safe and unsafe conditioned teacher student distributions are filtered and aggregated into a hindsight safety score, retaining both positive and negative evidence along student generated trajectories. We organize prefix branched rollouts of the same task into a distillation tree, where actions sharing the same interaction history form sibling branches and different action orders correspond to distinct paths. Starting from the leaves, hindsight scores are recursively propagated through downstream descendants and normalized into branch beliefs. Following group-relative comparison, each turn receives credit relative to its siblings, allowing the same action to receive different credit depending on its history, position, and future consequences. This formulation reuses existing rollouts without additional sampling or a separately trained value model or critic.

Our main contributions are as follows:
(i) Two-sided turn-level evidence. We use successful and failed privileged contexts to estimate each action's safety contribution, preserving discriminative credit signals as the policy improves.
(ii) Hindsight and Foresight Distillation Tree. We propagate local evidence over shared prefixes and contrast sibling branches to capture both interaction history and downstream consequences.
(iii) Empirical validation. Experiments on three safety benchmarks and two backbone models show that HFDTree improves safe task completion and counterfactual credit estimation while reducing unsafe actions and unnecessary refusals.

2 METHODOLOGY

We perform turn-level credit assignment for agentic value alignment on a *turn level hindsight and foresight tree* , a prefix tree that merges a group of sampled trajectories so that alternative continuations of the same interaction history become sibling branches. Each turn is scored by contrasting the likelihood of its action under success and failure related privileged contexts, measuring how strongly the action shifts the belief that the trajectory will satisfy the target value.

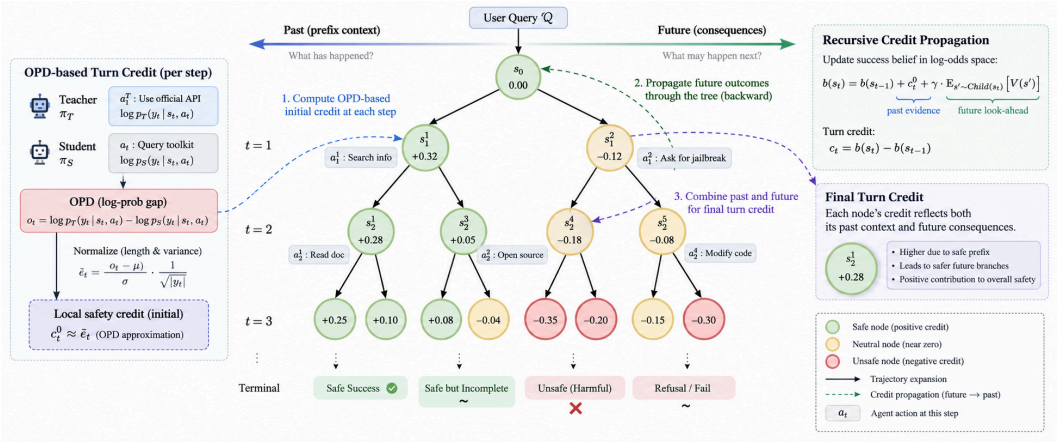

Figure 2: Overview of **HFDTree**. OPD-derived local turn credits are propagated through a prefix-conditioned turn level context consequence tree , allowing each action to be evaluated jointly from its interaction history and downstream consequences.

2.1 Problem Setup

Given a task x and an initial environment observation o_0, the agent starts from the state $s_1 = (x, o_0)$. At turn k it samples an action $a_k = (y_{k,1}, \ldots, y_{k,L_k}) \sim \pi_\theta(\cdot \mid s_k)$ from the current policy π_θ, where s_k is the observable interaction history so far, $y_{k,t}$ its t-th token, and L_k its length. Executing a_k yields environment feedback o_k and updates the history to $s_{k+1} = (s_k, a_k, o_k)$. A complete K-turn trajectory is $\tau = (s_1, a_1, o_1, \ldots, s_K, a_K, o_K)$. For each task x, we construct rollouts by branching from shared interaction prefixes rather than sampling independent trajectories. Selected prefixes are restored and expanded with multiple policy samples, ensuring explicit sibling branches at intermediate depths. Let C denote the event that a trajectory eventually satisfies the target value or safety constraint. For every realized turn (s_k, a_k), we estimate the contribution of a_k to C conditioned on the preceding interaction prefix, and thereby identify which intermediate decisions move the trajectory toward or away from the safe region.

2.2 From Outcome Contribution to Two-Sided Bayesian Evidence

We credit each turn by how strongly it drives the trajectory toward the target behavior. For an action a taken under the interaction history s, Bayes' rule gives under policy μ

$$\operatorname{logit} p_\mu(C \mid s, a) - \operatorname{logit} p_\mu(C \mid s) = \log \frac{p_\mu(a \mid s, C)}{p_\mu(a \mid s, \neg C)} =: \mathcal{B}(s, a), \tag{1}$$

where $\operatorname{logit}(u) = \log \frac{u}{1-u}$. Here $\mathcal{B}(s, a)$ is the ideal Bayesian evidence carried by the action, its sign indicating whether the action increases or decreases the odds of C. Conditioning on s makes the evidence history dependent, so superficially identical actions taken after different prefixes can contribute very differently.

Computing $\mathcal{B}(s, a)$ exactly is intractable because both likelihoods require marginalizing over all future continuations. We therefore approximate them via two-sided privileged self-distillation. For the current action, the same backbone evaluates it under a safe trajectory c^+ and a corresponding violating trajectory c^-, respectively. The resulting token-level contrastive signals are aggregated at the turn level, enabling fine-grained credit assignment while preserving the action taken under the original prefix. Writing $h_t^\pm = (s, c^\pm, y_{<t})$ for $a = (y_1, \ldots, y_L)$, the *two-sided evidence* is

$$e(s, a) = \sum_{t=1}^{L} \Big[\log \pi_T(y_t \mid h_t^+) - \log \pi_S(y_t \mid h_t^-) \Big] = \log \frac{\pi_T(a \mid s, c^+)}{\pi_S(a \mid s, c^-)}, \tag{2}$$

where π_T and π_S denote the same model π_ϕ conditioned on the positive and negative reference trajectories, respectively. We adopt the TCSD formulation in Eq. **??**, which keeps the candidate

action fixed and evaluates its token likelihood under the safe and unsafe full-trajectory references, respectively. which approximates Eq. 1 by comparing how compatible one and the same realized action is with a success- and with a failure-related context. The π_T do not have hindsight leakage.

Summing token-level log-likelihood differences makes $e(s,a)$ scale with action length even at constant average token evidence, so we normalize, $\bar{e}(s,a) = e(s,a)/L$, and suppress near-zero fluctuations with a sign-preserving, bounded soft-threshold, $\sigma_{T,\delta}(r) = \mathrm{sign}(r)\big[2\sigma((|r|-\delta)_+/T) - 1\big]$ and $\tilde{e}(s,a) = \sigma_{T,\delta}\big(\bar{e}(s,a)\big)$, where $\sigma(z) = 1/(1+\exp(-z))$, $(u)_+ = \max(u,0)$, $T > 0$ controls smoothness, and $\delta \geq 0$ sets a dead zone for weak evidence, beyond which the magnitude grows smoothly toward 1. Thus $\tilde{e}(s,a)$ retains whether the action supports or opposes the target value while bounding the influence of action length and weak fluctuations.

Proposition 2.1 (**Properties of Two-Sided Safety Evidence**). *Let $C, \neg C$ denote safety constraint satisfaction. For interaction state s and policy μ, suppose that scoring model π_ϕ, evaluated under positive and negative privileged contexts, approximates the corresponding outcome-conditioned action distributions: $|\log \pi_\phi(a \mid s, c^+) - \log p_\mu(a \mid s, C)| \leq \epsilon_+$, and $|\log \pi_\phi(a \mid s, c^-) - \log p_\mu(a \mid s, \neg C)| \leq \epsilon_-$. Define the token-normalized two-sided evidence as $\bar{e}_\phi(s,a) = L^{-1}\log[\pi_\phi(a \mid s, c^+)/\pi_\phi(a \mid s, c^-)]$.*
Then it has three properties: ***(i) Approximate Bayesian Evidence:*** *$\bar{e}_\phi$ estimates the per-token Bayesian update $L^{-1}[\mathrm{logit}\, p_\mu(C \mid s,a) - \mathrm{logit}\, p_\mu(C \mid s)]$ within error $(\epsilon_+ + \epsilon_-)/L$.*
(ii) Prior-Invariant Discrimination: *Unlike the one-sided contrast $\log \frac{p_\mu(a|s,C)}{p_\mu(a|s)}$, which degenerates as $p_\mu(C \mid s) \to 1$, the two-sided contrast is independent of the success prior; under exact scoring and fixed L, its expected class separation equals the per-token Jeffreys divergence.*
(iii) Robust Bounded Sensitivity: *Let $\tilde{e} = \psi_{T,\delta}(\bar{e}_\phi)$, where $\psi_{T,\delta}(r) = \mathrm{sign}(r)\tanh((|r|-\delta)_+/2T)$. This transformation is bounded by $|\tilde{e}| \leq 1$ and is $1/(2T)$-Lipschitz. Thus, length normalization removes linear dependence on token count, while thresholding suppresses weak evidence and bounds sensitivity to token-level log-likelihood perturbations.*

This formulation avoids the low-success-rate approximation, preserves discrimination as success rates increase, and improves robustness to action length and small likelihood perturbations. The resulting $\tilde{e}(s,a)$ is computed once per distinct state action pair and reused as local evidence in the subsequent context consequence tree. Detailed proof of the theorem and empirical validation of its effectiveness are provided in the appendix B and appendix B.1

2.3 Building the context consequence tree from turn Level Evidence

While $\tilde{e}(s,a)$ captures local safety evidence under the observed history, an action's safety consequences also depend on subsequent interactions. We therefore organize the sampled trajectories into a turn that folds downstream evidence into turn-level credit.

Given $\mathcal{G}_x = \{\tau_1, \ldots, \tau_G\}$, we build a rooted tree $\mathcal{T}_x = (\mathcal{N}_x, \mathcal{E}_x)$ whose root is the shared initial state $s_r = (x, o_0)$. The tree is constructed directly from prefix-branching rollouts rather than by post-hoc trajectory matching. Each node represents an exact shared prefix, from which multiple actions are sampled to form sibling branches; no semantic or approximate merging is used. executing a_n under s_p and receiving o_n creates a child n with history $s_n = (s_p, a_n, o_n)$, and the edge $p \to n$ carries the local evidence $\tilde{e}_{p\to n} = \tilde{e}(s_p, a_n)$. Sibling branches thus compare alternative actions and their observed continuations under a common prefix, controlling for earlier interaction history.

The local evidence $\tilde{e}(s_p, a_n)$ captures only the immediate safety signal, so we propagate downstream information through the turn-level context-consequence tree. For each non-root node n, with parent $p(n)$ and local evidence $\tilde{e}_n = \tilde{e}(s_{p(n)}, a_n)$, we estimate empirical transition probabilities $\hat{P}(m \mid n)$ from the sampled continuations and recursively compute the discounted downstream evidence U_n, effective horizon Z_n, and normalized branch belief B_n:

$$U_n = \tilde{e}_n + \gamma \sum_{m \in \mathrm{ch}(n)} \hat{P}(m \mid n)\, U_m, \qquad Z_n = 1 + \gamma \sum_{m \in \mathrm{ch}(n)} \hat{P}(m \mid n)\, Z_m, \qquad B_n = \frac{U_n}{Z_n}, \quad (3)$$

where $\gamma \in [0,1]$ controls how strongly future evidence influences the current decision; at a leaf the recursion degenerates to $U_n = \tilde{e}_n$ and $Z_n = 1$. Thus B_n is the normalized safety belief of the entire branch that follows the action into n: $B_n > 0$ indicates that the action together with its continuations

is on the whole value-aligned, $B_n < 0$ that the branch has accumulated more risk evidence. Two actions with nearly identical $\tilde{e}_n$ can therefore receive very different B_n if one repeatedly enters risky branches afterwards while the other stays safe, and a locally positive action can be strongly attenuated by downstream risk. We additionally include experiments on the depth and width of the tree under the above setting in the appendix.

Turn-level credit. B_n describes the whole branch that starts at the current action; we now ask how much choosing that action, under a fixed prefix, shifts the downstream belief. With the empirical mean belief over sibling branches $\bar{B}_p = \sum_{j \in \mathrm{ch}(p)} \hat{P}(j \mid p)\, B_j$, the turn-level credit is

$$\Delta_n = \begin{cases} B_n - \bar{B}_{p(n)}, & |\,\mathrm{ch}(p(n))| > 1, \\ B_n, & |\,\mathrm{ch}(p(n))| = 1. \end{cases} \tag{4}$$

When several continuations are observed under the same prefix, Δ_n measures the deviation of the downstream belief after choosing branch n from the average of the alternatives available there, and hence satisfies $\sum_{n \in \mathrm{ch}(p)} \hat{P}(n \mid p)\, \Delta_n = 0$. Finally, for each task x, we use a correct trajectory provided by the dataset as the safe reference, $\tau^+ = (s_1^+, a_1^+, o_1^+, \ldots, s_K^+, a_K^+, o_K^+)$. that satisfies the target safety constraint; when the teacher cannot provide reliable behavioral guidance for a turn, the corresponding safe step (s_k^+, a_k^+) serves as a direct supervision signal.

2.4 Safety Credit Guided Policy Optimization

We use the clipped turn-level credit $A_n = \mathrm{clip}(\Delta_n, -1, 1)$ as a signed learning signal, reinforcing positive credit actions and suppressing negative-credit ones. When both local evidence and propagated credit are weak, we instead use a state compatible safe reference action. Specifically, fallback supervision is applied only to nodes admitting a valid reference match, $\mathcal{R} = \{n \in \mathcal{N} \cap \mathcal{V} \mid |\bar{e}_n| \le \delta \wedge |\Delta_n| \le \epsilon\}$, where $\mathcal{V}$ contains nodes whose matched reference action is executable in the current environment state. Let $\ell_\theta(n)$ and $\ell_\theta^+(n)$ denote the average token log-likelihoods of the current action and the matched safe reference action, respectively, both conditioned on the current rollout prefix:

$$\ell_\theta(n) = \frac{1}{L_n} \sum_{t=1}^{L_n} \log \pi_\theta\big(y_{n,t} \mid s_{p(n)}, y_{n,<t}\big), \qquad \ell_\theta^+(n) = \frac{1}{L_{m(n)}^+} \sum_{t=1}^{L_{m(n)}^+} \log \pi_\theta\big(y_{m(n),t}^+ \mid s_{p(n)}, y_{m(n),<t}^+\big). \tag{5}$$

The training objective then selects, per turn, between credit-guided optimization and reference supervision:

$$\mathcal{L} = \frac{1}{|\mathcal{N}|} \sum_{n \in \mathcal{N}} \mathcal{L}_n, \qquad \mathcal{L}_n = \begin{cases} -\,\mathrm{sg}(A_n)\, \ell_\theta(n), & n \notin \mathcal{R}, \\ -\ell_\theta^+(n), & n \in \mathcal{R}, \end{cases} \tag{6}$$

where $\mathrm{sg}(\cdot)$ denotes the stop gradient operator. Outside $\mathcal{R}$, we optimize with tree based credit; inside $\mathcal{R}$, we supervise with a state-valid safe reference action. If no valid reference match exists, the fallback is omitted.

Computational complexity and practical setup. Tree construction takes $O(GK)$ time, while propagation requires $O(N)$ time and memory; the main additional cost is scoring action tokens under the two privileged contexts.

3 Experiments

We ask whether HFDTree improves agent safety without sacrificing task utility, and whether the improvement is attributable to accurate turn level credit. Our evaluation covers overall safety utility performance (§3.2), robustness to long horizons, sparse critical decisions and delayed consequences (§3.4), the validity of the estimated credit (§3.3), and sensitivity to safety relevant action ordering (§3.5).

Table 1: **Overall safety–utility comparison across agent-safety benchmarks and model backbones.** Baselines are grouped by the component of our method they isolate. We report SVR (↓), ORR (↓), and SCR (↑); all methods use comparable training settings within each backbone, and results are mean ± standard deviation over three runs. **Bold** denotes the best result, and underlined the strongest prior method.

Backbone	Method	AgentHarm			Agent-SafetyBench			SafeAgentBench		
		SVR↓	ORR↓	SCR↑	SVR↓	ORR↓	SCR↑	SVR↓	ORR↓	SCR↑
Qwen3-4B	*Credit-free controls*									
	Base Model	36.8±0.9	8.1±0.5	49.6±1.0	54.4±1.1	9.7±0.6	28.5±1.1	42.1±1.2	8.9±0.6	11.8±0.7
	safe RLHF-v	21.6±0.8	9.1±0.6	57.3±0.8	40.6±0.9	9.3±0.5	38.1±0.9	28.1±1.0	8.6±0.5	25.0±0.8
	Privileged-information credit									
	OPPO	21.5±0.4	10.2±0.6	49.3±0.7	37.5±0.8	9.9±0.5	43.4±0.3	21.9±0.3	10.3±0.4	31.2±0.7
	PBSD	17.5±0.8	8.2±0.5	59.4±0.8	36.8±0.8	8.0±0.5	42.9±0.8	22.4±0.9	8.1±0.5	30.9±0.8
	OVCSD	26.4±1.0	10.7±0.7	53.2±0.9	46.8±1.0	11.1±0.7	33.4±1.0	35.7±1.2	10.6±0.7	17.4±0.8
	OPSD	16.1±0.7	7.3±0.4	60.8±0.7	33.6±0.8	7.2±0.4	46.0±0.8	18.7±0.8	7.0±0.4	35.4±0.7
	Turn-level credit assignment									
	GraphGPO	19.8±0.9	8.5±0.5	58.2±0.8	38.2±0.9	8.4±0.5	41.3±0.9	24.6±1.0	8.3±0.5	28.8±0.8
	AT^{2}PO	14.2±0.7	7.1±0.4	62.3±0.7	31.9±0.7	7.0±0.4	47.6±0.7	17.5±0.8	6.8±0.4	36.2±0.7
	Agent safety alignment									
	AgentAlign	11.2±0.7	7.8±0.5	61.7±0.7	36.0±0.8	7.6±0.5	44.1±0.8	21.0±0.9	7.7±0.4	32.7±0.8
	SafeBranch	13.5±0.7	7.5±0.4	61.0±0.7	35.2±0.8	7.5±0.4	44.8±0.8	16.9±0.8	7.2±0.4	36.8±0.7
	HFDTree (Ours)	**8.9**±0.5	**6.1**±0.3	**64.7**±0.6	**29.4**±0.6	**6.2**±0.3	**51.4**±0.7	**13.8**±0.7	**6.1**±0.3	**41.9**±0.7
LLaMA-3.1-8B-Instruct	*Credit-free controls*									
	Base Model	34.6±0.3	8.6±0.6	46.3±0.8	55.7±0.3	10.3±0.4	26.1±0.7	43.6±0.9	9.4±0.4	10.7±0.2
	safe RLHF-v	20.9±0.9	9.6±0.7	54.5±0.4	41.4±0.5	9.9±0.5	35.5±0.7	29.7±0.9	9.1±0.5	22.7±0.3
	Privileged-information credit									
	OPPO	22.5±0.5	12.2±0.8	50.3±0.4	36.3±0.5	10.1±0.8	42.8±0.6	22.3±0.9	11.7±0.5	33.2±0.6
	PBSD	16.8±0.6	8.6±0.7	57.2±0.5	37.9±0.8	8.5±0.9	40.4±0.4	23.9±0.5	8.5±0.6	28.7±0.8
	OVCSD	25.0±1.0	11.2±0.7	50.1±0.6	48.6±0.8	11.8±1.1	30.8±1.0	37.3±1.2	11.2±0.7	16.0±0.4
	OPSD	15.4±0.5	7.7±0.8	58.7±0.3	35.1±0.8	7.6±0.4	43.6±0.5	20.1±0.7	7.4±0.9	32.8±0.3
	Turn-level credit assignment									
	GraphGPO	19.1±0.6	8.9±0.8	55.9±0.2	39.4±0.3	8.8±0.4	38.9±0.5	26.1±1.1	8.7±0.1	26.5±0.9
	AT^{2}PO	13.6±0.7	7.5±0.6	60.2±0.8	33.4±0.3	7.4±0.6	45.1±0.7	18.9±0.9	7.2±0.5	33.6±0.6
	Agent safety alignment									
	AgentAlign	10.5±0.5	8.2±0.6	59.7±0.8	37.1±0.8	8.1±0.3	41.7±0.4	22.6±0.9	8.0±0.4	30.6±0.8
	SafeBranch	12.8±0.6	7.9±0.7	58.9±0.5	36.4±0.8	8.0±0.4	42.3±0.8	18.3±0.9	7.6±0.4	34.5±0.9
	HFDTree (Ours)	**8.1**±0.7	**6.5**±0.8	**62.6**±0.4	**30.9**±0.5	**6.5**±0.3	**48.5**±0.7	**15.0**±0.8	**6.4**±0.3	**38.7**±1.1

3.1 Experimental Setup

Benchmarks and backbones. We evaluate on AgentHarm (Andriushchenko et al., 2025), Agent-SafetyBench (Zhang et al., 2024), and SafeAgentBench (Yin et al., 2024), covering request handling, interactive safety, and safe planning and execution. We use Qwen3-4B and LLaMA-3.1-8B-Instruct as the backbone models for the bidirectional-information self-distillation tree. (Appendices C.1 and C.3).

Baselines and controls. We compare with the unaligned base model, outcome-supervised Safe RLHF-v (Ji et al., 2026a), and three baseline groups: (i) *privileged-information credit*, OPPO (Li et al., 2026c), OPSD (Zhao et al., 2026a), OVCSD (Xia et al., 2026) and PBSD (Tian et al., 2026); (ii) *turn-level credit*, GraphGPO (Cheng et al., 2026), and AT^{2}PO (Zong et al., 2026); and (iii) *agent safety alignment*, AgentAlign (Zhang et al., 2025) and SafeBranch (Lee et al., 2026). For directly comparable methods, we fix the student backbone, task split, rollout budget, and terminal safety feedback, and match teacher configuration and supervision where applicable (Appendix C.2).

Metrics. We report safety violation rate (SVR↓), over-refusal rate (ORR↓), and safe completion rate (SCR↑) to jointly assess safety and utility. On SafeAgentBench, we also compare methods at matched benign or harmful task success rates (Appendix C.4).

Diagnostic protocols. Using Qwen3-4B on SafeAgentBench extensions, we evaluate robustness to horizon K, critical decision count M, and consequence delay D, with $K = 32$ fixed to isolate decision sparsity. Bayesian evidence credit is estimated from repeated continuations of original and alternative actions under a fixed prefix and policy, and cross checked by an independent LLM judge. Order tests include safety irrelevant swaps as controls, while all ablations use the same training

budget (Appendices C.1 and C.4). More reproducibility details of the experimental setup are provided in the appendix C.

3.2 Main Results

HFDTree consistently improves both safety and task utility. As shown in Table 1, HFDTree achieves the lowest SVR and ORR and the highest SCR across all three benchmarks and both backbones. On Qwen3-4B, it reduces SVR by 2.3–3.1 percentage points and improves SCR by 2.4–5.1 percentage points relative to the strongest baseline for each metric. The largest SCR gain occurs on SafeAgentBench, highlighting improved task completion under safety constraints. These gains are accompanied by lower over refusal across all evaluated settings, indicating that HFDTree improves safety while remaining responsive to benign requests. This joint improvement is consistent with HFDTree's approach to assigning credit to individual decisions. Two sided evidence assesses each action under success and failure related contexts, while HFDTree incorporates downstream safety evidence and compares alternative branches under the same interaction history. Together, these mechanisms provide a training signal designed to discourage actions that lead to unsafe continuations while reinforcing those that support safe task completion.

Table 2: **General capability preservation after safety alignment.** We compare Qwen3-4B before and after safety alignment on general agent benchmarks.

	Safety		General Agent Capability		
Method	SVR↓	SCR↑	BFCL-ST↑	BFCL-MT↑	τ-bench↑
Qwen3-4B	42.1	11.8	**57.6**	46.3	**51.2**
AT^{2}PO	17.5	36.2	43.8	35.9	45.2
HFDTree	**13.8**	**41.9**	57.3	**47.6**	50.9

HFDTree preserves general agent capability after safety alignment. Table 2 compares the original Qwen3-4B with safety aligned models on general agent capability benchmarks. HFDTree substantially improves safety performance while largely preserving general tool use and task solving capability, indicating limited alignment tax. Detailed safety utility trade-off results, evaluation protocols, and benchmark descriptions are provided in the appendix.

3.3 Validity of Turn Level Credit

HFDTree prioritizes turns with substantial effects on downstream safety. At a 20% inspection budget, the full credit Δ_n captures 62% of the absolute Bayesian evidence contribution (Figure 3a). Local two sided evidence captures 59%, improving upon one sided evidence by 15 percentage points, while tree correction adds a further 3 percentage points. These results support the success and failure conditioned contrast in Eq. 2 for identifying consequential decisions.

HFDTree tracks how an action's contribution changes with its preceding context. Figure 3b compares the same action under paired prefixes while holding the continuation policy fixed. The full credit achieves a Spearman correlation of 0.98 and 100% sign agreement with environment measured contribution shifts, compared with 0.93 and 93% for local two sided evidence. The corresponding mean absolute error decreases from 0.13 to 0.05. Local evidence already captures prefix dependence, while tree correction improves its correspondence with the Bayesian evidence changes. Neutral prefix interventions provide a control for sensitivity to safety irrelevant changes.

Tree correction improves attribution for decisions with delayed safety consequences. Across the ten cases in Figure 3c, with delays ranging from 2 to 16 turns, the full credit has lower attribution error than both local two sided evidence. Its mean error is 0.03, compared with 0.27 and 0.34, respectively. These case level results support incorporating downstream evidence when local scores do not adequately reflect an earlier decision's safety consequences. We estimate Bayesian evidence safety advantage against shared alternative actions using repeated continuations independent of credit construction. We evaluate critical turn capture, prefix shift agreement, and delayed attribution error, with MAE computed on a common scale fixed without test set fitting. Appendix D.3 provides the protocol and metrics.

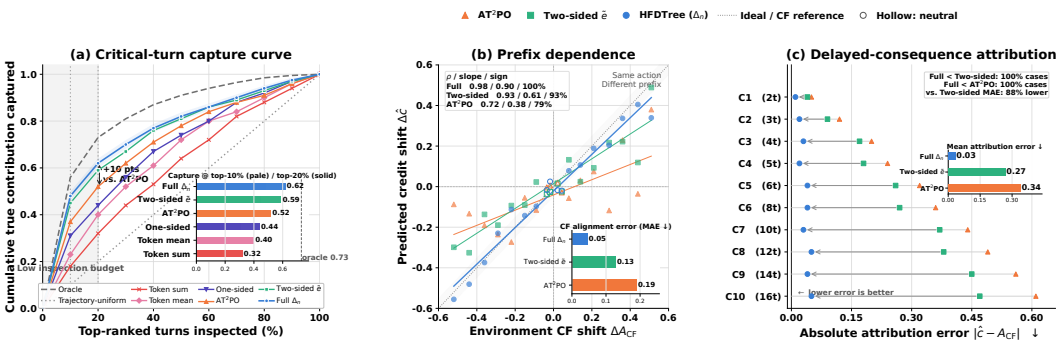

Figure 3: HFDTree identifies consequential turns and better aligns credit with Bayesian evidence safety effects. (a) Credit ranked turns capture most cumulative contribution. (b) Predicted credit shifts track Bayesian evidence changes across prefixes for the same action. (c) HFDTree reduces attribution error on delayed consequence cases; insets summarize the corresponding metrics.

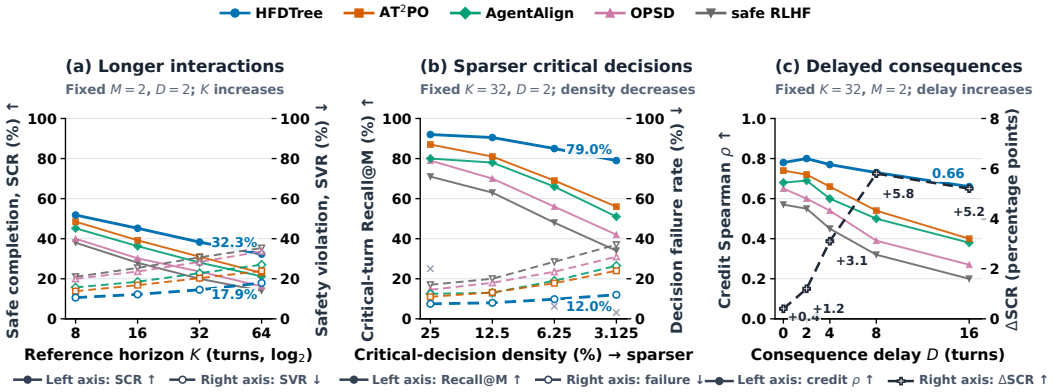

Figure 4: Robustness to (a) longer interaction horizons, (b) sparser critical decisions, and (c) delayed consequences. We fix $(M, D) = (2, 2)$ in (a), $(K, D) = (32, 2)$ in (b), and $(K, M) = (32, 2)$ in (c). Solid and dashed curves use the left and right axes, respectively. $\Delta\text{SCR} = \text{SCR}*\text{Full} - \text{SCR}*\gamma = 0$ denotes the gain from downstream propagation under the same training budget.

3.4 Robustness to Horizon, Sparsity, and Delay

Figure 4 shows that HFDTree remains robust to long horizons, sparse critical decisions, and delayed consequences. At $K = 64$, it retains a 32.3% SCR with a 17.9% SVR; when critical decisions constitute only 3.125% of all turns, it achieves 79.0% Recall@M with a 12.0% conditional failure rate. Moreover, at $D = 16$, HFDTree still improves SCR by 5.2 percentage points over the equally trained $\gamma = 0$ variant, while its credit maintains a $\rho = 0.66$ correlation with Bayesian evidence contributions. These results indicate that combining local turn level evidence with downstream propagation helps identify safety critical decisions and preserve safe task completion in long-horizon interactions.

3.5 Credit Construction and Agent Safety

Ablation design. Table 3 separates the effects of evidence construction, future propagation, and sibling centering. The one sided + HFDTree variant changes evidence construction while retaining propagation and centering. With two-sided evidence fixed, local only, future only, centering only, and full HFDTree form a 2×2 ablation of the two credit-assignment operations. Appendix D.5 details the evaluation protocol and metrics.

Two-sided evidence improves credit quality and policy safety. With both HFDTree operations retained, replacing one-sided evidence with two-sided evidence improves Bayesian evidence agree-

Table 3: **Ablation of credit construction and agent safety.** Metrics assess credit quality and Qwen3-4B safety on SafeAgentBench. *Bi-reference supervision* receives the same positive and negative reference information as HFDTree but uses it directly for supervision without tree-based credit assignment. Accuracies and rates are percentages; "–" denotes unavailable or inapplicable entries.

	Components				Credit Quality			Trained-Policy Safety		
Variant	Two-sided evidence	Future	Centering	Safe-ref. fallback	CF $\rho\uparrow$	Delayed Acc.↑	Order Acc.↑	SVR↓	ORR↓	SCR↑
AT^{2}PO	–	–	–	–	–	–	–	17.5 ± 0.8	6.9 ± 0.4	36.4 ± 0.7
Bi-reference supervision	✗	✗	✗	–	–	–	–	15.9 ± 0.7	6.7 ± 0.4	39.0 ± 0.7
Controlled component ablations										
One-sided + HFDTree	✗	✓	✓	✓	.61	71.3	73.8	16.6	6.6	38.2
Two-sided **local only**	✓	✗	✗	✓	.31	52.7	55.8	18.1	6.8	35.9
Two-sided + **future**	✓	✓	✗	✓	.67	74.2	70.4	15.7	6.5	39.6
Two-sided + **centering**	✓	✗	✓	✓	.43	56.9	67.1	17.2	6.4	37.4
Full HFDTree **w/o fallback**	✓	✓	✓	✗	.79	82.9	85.2	14.6	6.4	40.7
Full HFDTree	✓	✓	✓	✓	**.79**	**82.9**	**85.2**	$\mathbf{13.8} \pm 0.7$	$\mathbf{6.1} \pm 0.3$	$\mathbf{41.9} \pm 0.7$

ment, delayed-consequence attribution, and action-order accuracy. These gains are accompanied by fewer safety violations and higher safe completion, supporting the value of contrasting success- and failure-conditioned evidence for both credit estimation and policy learning. Propagation connects later safety consequences to earlier decisions, while centering expresses propagated evidence relative to alternatives under the same prefix. Their respective ablations show improved delayed-consequence attribution and action-order agreement, alongside better policy safety and completion.

4 Related Work

Agent safety alignment and fine-grained credit assignment. *Safety alignment* uses inference-time guards to monitor agent actions (Xiang et al., 2024; Wang et al., 2025), although isolated judgments can miss contextual risks (Chen et al., 2026). Parametric methods learn from safety demonstrations or trajectory-level rewards (Rafailov et al., 2023; Shao et al., 2024; Zhang et al., 2025), often without distinguishing safety-critical turns; SafeBranch (Lee et al., 2026) instead supervises safe/unsafe alternatives at violating steps. *Fine-grained credit assignment* localizes task rewards through critics, process supervision, turn-level objectives, or branching estimators (Zhou et al., 2025; Xi et al., 2026; Li et al., 2026a; Hong et al., 2026; Ding & Ye, 2026; Zhang & Stadie, 2026; Zong et al., 2026). HFDTree propagates downstream outcome beliefs and centers them on sibling averages to assign safety credit sensitive to action order.

Privileged-information distillation. Privileged teachers provide dense supervision (Vapnik & Izmailov, 2015; Lopez-Paz et al., 2015; Zhao et al., 2026a), including through successful hindsight trajectories (Yeo et al., 2026; Li et al., 2026b; Jin et al., 2026). PBSD (Tian et al., 2026) contrasts outcome-conditioned and marginal likelihoods, yielding a one-sided approximation that loses discrimination as success rates rise. Our two-sided approach contrasts contexts to approximate Bayesian evidence directly. Appendix A provides the full discussion.

5 Conclusion

We presented **HFDTree**, a framework for turn-level safety credit assignment in LLM agents. HFDTree combines two-sided privileged likelihood evidence with a hindsight foresight distillation tree: local action evidence is estimated under safe and unsafe contexts, then refined by propagating downstream consequences and comparing sibling branches under shared prefixes. This allows credit to reflect both prior interaction context and future outcomes, including action order effects. Experiments across three agent safety benchmarks and two backbone models show improved safety, safe task completion, and counterfactual credit estimation. Future work will study more efficient branch exploration and broader open-ended agent settings.

AI USE STATEMENT

Generative AI tools were used during manuscript preparation to assist with language editing, LaTeX formatting. All methodological decisions, implementations, experimental results, analyses, citations, and scientific claims were independently reviewed and verified by the authors. Generative AI tools were not used to autonomously conduct experiments, generate or modify reported measurements, or determine the conclusions of this work. The authors take full responsibility for the content of the paper.

ETHICS STATEMENT

This work studies safety alignment for LLM-based agents and is intended to reduce unsafe behavior during multi-steps interaction and tool use. Our experiments use existing research benchmarks and controlled evaluation environments; we do not conduct studies involving human subjects or collect private user data. Some benchmarks contain harmful requests or safety critical scenarios as part of their evaluation protocols. These examples are used solely to measure and improve model safety in controlled settings, and our experiments do not execute harmful actions in real-world systems. As with other work on agent safety, analyses of model failure modes may have dual use implications. We therefore focus on aggregate evaluation and mitigation methods rather than operational guidance for causing harm. We follow the licenses and terms of use of the models and benchmarks used in this study and adhere to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

We provide the methodological and experimental details needed to reproduce our results throughout the paper and appendices. Section 2 specifies the construction of the two-sided turn-level evidence, the Hindsight and Foresight Distillation Tree, its belief propagation and sibling-centered credit assignment, and the resulting policy objective. Section 3.1 summarizes the benchmarks, model backbones, baselines, and controlled training settings. Appendix C provides detailed descriptions of the benchmarks, baselines, backbone configurations, evaluation metrics, and controlled model settings, while Appendix D specifies the counterfactual credit-evaluation protocols, diagnostic experiments, and ablation settings. For directly comparable methods, we keep the student backbone, task split, rollout budget, decoding configuration, interaction budget, and available supervision fixed whenever applicable. Reported main results are averaged over three runs. We will provide the implementation and experiment configurations in the supplementary material to facilitate reproduction of the reported results.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, Zico Kolter, Matt Fredrikson, et al. Agentharm: A benchmark for measuring harmfulness of llm agents. In *International Conference on Learning Representations*, volume 2025, pp. 79185–79220, 2025.

Yen-Shan Chen, Sian-Yao Huang, Cheng-Lin Yang, and Yun-Nung Chen. Tracesafe: A systematic assessment of llm guardrails on multi-step tool-calling trajectories. *arXiv preprint arXiv:2604.07223*, 2026.

Xin Cheng, Shuo He, Lang Feng, HaiYang Xu, Ming Yan, Lei Feng, and Bo An. Beyond trajectory-level attribution: Graph-based credit assignment for agentic reinforcement learning. *arXiv preprint arXiv:2605.26684*, 2026.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Juntao Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. In *International Conference on Learning Representations*, volume 2024, pp. 50750–50777, 2024.

Zheng Ding and Weirui Ye. Treegrpo: tree-advantage grpo for online rl post-training of diffusion models. In *International Conference on Learning Representations*, volume 2026, pp. 10877–10889, 2026.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Zhepei Hong, Lin Wang, Liting Li, Haokai Ma, Junfeng Fang, Fei Shen, Dan Zhang, and Xiang Wang. Trace: Trajectory risk-aware compression for long-horizon agent safety. *arXiv preprint arXiv:2606.00611*, 2026.

Jiaming Ji, Xinyu Chen, Rui Pan, Han Zhu, Jiahao Li, Donghai Hong, Boyuan Chen, Jiayi Zhou, Kaile Wang, Juntao Dai, et al. Safe rlhf-v: Safe reinforcement learning from multi-modal human feedback. *Advances in Neural Information Processing Systems*, 38:46146–46182, 2026a.

Yuxiang Ji, Ziyu Ma, Yong Wang, Guanhua Chen, Xiangxiang Chu, and Liaoni Wu. Tree search for llm agent reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 87362–87388, 2026b.

Xiaolong Jin, Dingmin Wang, Vijay Lingam, and Varun Kumar. Ahead: Adaptive hindsight with environment-augmented distillation for agentic rl. *arXiv preprint arXiv:2608.24114*, 2026.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Ren Lu, Thomas Mesnard, Johan Ferret, Colton Bishop, Ethan Hall, Victor Carbune, and Abhinav Rastogi. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. 2023.

Hyunse Lee, Jiwoo Jeong, Haneul Lee, Kyochul Jang, Youngjae Yu, and Woojin Lee. Safebranch: Branch-pair safety alignment for embodied agents. *arXiv preprint arXiv:2608.19729*, 2026.

Junbo Li, Peng Zhou, Rui Meng, Meet P Vadera, Lihong Li, and Yang Li. Turn-ppo: Turn-level advantage estimation with ppo for improved multi-turn rl in agentic llms. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 6227–6243, 2026a.

Xiaozhe Li, Tianyi Lyu, Yang Li, Yichuan Ma, Peiji Li, Linyang Li, Qipeng Guo, Dahua Lin, and Kai Chen. What and when to distill: Selective hindsight distillation for multi-turn agents. *arXiv preprint arXiv:2605.19447*, 2026b.

Yu Li, Rui Miao, Tian Lan, and Zhengling Qi. Oppo: Bayesian value recursion for token-level credit assignment in llm reasoning. *arXiv preprint arXiv:2605.21851*, 2026c.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. In *International Conference on Learning Representations*, volume 2024, pp. 52989–53046, 2024.

Yue Liu, Hongcheng Gao, Shengfang Zhai, Yufei He, Jun Xia, Zhengyu Hu, Yulin Chen, Xihong Yang, Jiaheng Zhang, Stan Z Li, et al. Guardreasoner: Towards reasoning-based llm safeguards. *arXiv preprint arXiv:2501.18492*, 2025.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015.

Weidi Luo, Shenghong Dai, Xiaogeng Liu, Suman Banerjee, Huan Sun, Muhao Chen, and Chaowei Xiao. Agrail: A lifelong agent guardrail with effective and adaptive safety detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8104–8139, 2025.

Aengus Lynch, Benjamin Wright, Caleb Larson, Stuart J Ritchie, Soren Mindermann, Evan Hubinger, Ethan Perez, and Kevin Troy. Agentic misalignment: How llms could be insider threats. *arXiv preprint arXiv:2510.05179*, 2025.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

Jian Mu, Qixin Zhang, Zhiyong Wang, Menglin Yang, Shuang Qiu, Chengwei Qin, Zhongxiang Dai, and Yao Shu. Self-reflective generation at test time. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10221–10239, 2026.

Xuan-Phi Nguyen, Shrey Pandit, Yiran Zhao, Anurag Koul, Zeyu Liu, and Shafiq Joty. Privileged likelihood is not automatically value: Three checks for token credit in on-policy self-distillation. *arXiv preprint arXiv:2608.09263*, 2026.

Shishir G Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E Gonzalez. The berkeley function calling leaderboard (bfcl): From tool use to agentic evaluation of large language models. In *Forty-second International Conference on Machine Learning*, 2025.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris Maddison, and Tatsunori Hashimoto. Identifying the risks of lm agents with an lm-emulated sandbox. In *International Conference on Learning Representations*, volume 2024, pp. 27031–27098, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Hang Su, Jun Luo, Chang Liu, Xiao Yang, Yichi Zhang, Yinpeng Dong, and Jun Zhu. A survey on autonomy-induced security risks in large model-based agents. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Kimi Team, Yifan Bai, Yiping Bao, Y Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

Meituan LongCat Team, Anchun Gui, Bei Li, Bingyang Tao, Bole Zhou, Borun Chen, Chao Zhang, Chen Gao, Chen Zhang, Chengcheng Han, et al. Longcat-flash-thinking-2601 technical report. *arXiv preprint arXiv:2601.16725*, 2026.

Yang Tian, Rui Wang, Xumeng Wen, Junjie Li, Shizhao Sun, Lei Song, Jiang Bian, and Bo Zhao. Pbsd: Privileged bayesian self-distillation for long-horizon credit assignment. *arXiv preprint arXiv:2606.09348*, 2026.

Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: similarity control and knowledge transfer. *The Journal of Machine Learning Research*, 16(1):2023–2049, 2015.

Haoyu Wang, Christopher M Poskitt, and Jun Sun. Agentspec: Customizable runtime enforcement for safe and reliable llm agents. *arXiv preprint arXiv:2503.18666*, 2025.

Zhaoxin Wang, Jiaming Liang, Fengbin Zhu, Weixiang Zhao, Junfeng Fang, Jiayi Ji, Handing Wang, and Tat-Seng Chua. Safeneuron: Neuron-level safety alignment for large language models. *arXiv preprint arXiv:2602.12158*, 2026.

Quan Wei, Siliang Zeng, Chenliang Li, William Brown, Oana Frunza, Wei Deng, Anderson Schneider, Yuriy Nevmyvaka, Yang Katie Zhao, Alfredo Garcia, et al. Reinforcing multi-turn reasoning in llm agents via turn-level reward design. *arXiv preprint arXiv:2505.11821*, 2025.

Zhiheng Xi, Chenyang Liao, Guanyu Li, Zhihao Zhang, Wenxiang Chen, Binghai Wang, Senjie Jin, Yuhao Zhou, Jian Guan, Wei Wu, et al. Agentprm: Process reward models for llm agents via step-wise promise and progress. In *Proceedings of the ACM Web Conference 2026*, pp. 4184–4195, 2026.

Xu Xia, Jinghua Piao, Min Yang, Xiaochong Lan, Jiaju Chen, and Yong Li. From scoring to acting: Outcome-verified comparative self-distillation for llm agents. *arXiv preprint arXiv:2607.27937*, 2026.

Zhen Xiang, Linzhi Zheng, Yanjie Li, Junyuan Hong, Qinbin Li, Han Xie, Jiawei Zhang, Zidi Xiong, Chulin Xie, Carl Yang, et al. Guardagent: Safeguard llm agents by a guard agent via knowledge-enabled reasoning. *arXiv preprint arXiv:2406.09187*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. tau-bench: A benchmark for\underline {T} ool-\underline {A} gent-\underline {U} ser interaction in real-world domains. In *International Conference on Learning Representations*, volume 2025, pp. 9965–10017, 2025.

Woongyeng Yeo, Yumin Choi, Taekyung Ki, and Sung Ju Hwang. Hint-sd: Targeted hindsight self-distillation for long-horizon agents. *arXiv preprint arXiv:2605.17873*, 2026.

Sheng Yin, Xianghe Pang, Yuanzhuo Ding, Menglan Chen, Yutong Bi, Yichen Xiong, Wenhao Huang, Zhen Xiang, Jing Shao, and Siheng Chen. Safeagentbench: A benchmark for safe task planning of embodied llm agents. *arXiv preprint arXiv:2412.13178*, 2024.

Ashkan Yousefpour, Taeheon Kim, Ryan S Kwon, Seungbeen Lee, Wonje Jeung, Seungju Han, Alvin Wan, Harrison Ngan, Youngjae Yu, and Jonghyun Choi. Representation bending for large language model safety. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24073–24098, 2025.

Tongxin Yuan, Zhiwei He, Lingzhong Dong, Yiming Wang, Ruijie Zhao, Tian Xia, Lizhen Xu, Binglin Zhou, Fangqi Li, Zhuosheng Zhang, et al. R-judge: Benchmarking safety risk awareness for llm agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1467–1490, 2024.

Chenchen Zhang. From reasoning to agentic: Credit assignment in reinforcement learning for large language models. *arXiv preprint arXiv:2604.09459*, 2026.

Jinchuan Zhang, Lu Yin, Yan Zhou, and Songlin Hu. Agentalign: Navigating safety alignment in the shift from informative to agentic large language models. *arXiv preprint arXiv:2505.23020*, 2025.

Zeyu Zhang and Bradly C Stadie. Tempo: Temporal enforcement via mode-separated policy optimization for trustworthy llm backtesting. *arXiv preprint arXiv:2605.18843*, 2026.

Zhexin Zhang, Shiyao Cui, Yida Lu, Jingzhuo Zhou, Junxiao Yang, Hongning Wang, and Minlie Huang. Agent-safetybench: Evaluating the safety of llm agents. *arXiv preprint arXiv:2412.14470*, 2024.

Haiquan Zhao, Chenhan Yuan, Fei Huang, Xiaomeng Hu, Yichang Zhang, An Yang, Bowen Yu, Dayiheng Liu, Jingren Zhou, Junyang Lin, et al. Qwen3guard technical report. *arXiv preprint arXiv:2510.14276*, 2025.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models, 2026a. URL `https://arxiv.org/abs/2601.18734`.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026b.

Yifei Zhou, Song Jiang, Yuandong Tian, Jason Weston, Sergey Levine, Sainbayar Sukhbaatar, and Xian Li. Sweet-rl: Training multi-turn llm agents on collaborative reasoning tasks. *arXiv preprint arXiv:2503.15478*, 2025.

Zefang Zong, Dingwei Chen, Yang Li, Qi Yi, Bo Zhou, Chengming Li, Bo Qian, Peng Chen, and Jie Jiang. At^{2}po: Agentic turn-based policy optimization via tree search. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24120–24143, 2026.

APPENDIX

This appendix expands the concise presentation in the main paper. It first positions HFDTree relative to the closest lines of work, then documents the experimental setup and evaluation criteria, and finally provides additional results and diagnostic protocols.

A EXTENDED RELATED WORK

Safety alignment for LLM agents. As summarized in Section 4, agent-safety methods either intervene during execution or update model parameters. Inference-time methods use reviewers, runtime monitors, or self-correction to block an action before it is executed (Xiang et al., 2024; Wang et al., 2025). A decision made from the current output alone, however, can miss risks that arise only from the preceding context or a later consequence. Parameter update methods instead train on safety demonstrations or terminal safety rewards (Rafailov et al., 2023; Shao et al., 2024; Zhang et al., 2025). A single trajectory-level target gives every turn the same supervision and therefore obscures the small number of decisions that determine safety. SafeBranch (Lee et al., 2026) narrows supervision to a safe/unsafe action pair at a violating step. HFDTree differs by comparing sibling actions through the downstream outcomes observed after each branch.

Privileged-information distillation. Privileged-information distillation obtains dense training signals from a teacher that observes information hidden from the student (Vapnik & Izmailov, 2015; Lopez-Paz et al., 2015; Zhao et al., 2026a). Hindsight methods commonly provide the teacher with a completed successful trajectory (Yeo et al., 2026; Li et al., 2026b; Jin et al., 2026). PBSD (Tian et al., 2026), for example, treats the likelihood gap between an outcome-conditioned teacher and the marginal policy as turn-level Bayesian evidence.

That construction replaces the failure-conditioned likelihood with a marginal likelihood under a low-success-rate approximation. Its one-sided contrast can therefore become less informative as the student succeeds more frequently. In addition, a likelihood gap derived from privileged information is not by itself a valid estimate of decision credit (Nguyen et al., 2026). Equation 2 avoids the substitution by evaluating the same realized action under both success and failure related contexts. The resulting contrast remains informative across different success rates and more closely follows the ideal Bayesian evidence. Section D.4 examines this interpretation empirically.

Turn-level and tree-structured credit assignment. Agentic reinforcement learning has addressed sparse outcome rewards with learned critics and process reward models (Zhou et al., 2025; Xi et al., 2026), turn-level objectives (Li et al., 2026a), and estimators defined over shared-prefix trees or branches (Ding & Ye, 2026; Zhang & Stadie, 2026; Zong et al., 2026). These methods typically target task completion and evaluate a turn from its prefix, its downstream return, or both.

HFDTree combines future evidence with a comparison among actions available under the same prefix. Belief propagation first assigns downstream outcomes to the branch-opening action; sibling centering then expresses that propagated belief relative to its alternatives. Centering after propagation distinguishes HFDTree from related shared-prefix estimators (Zong et al., 2026). Because a change in action order changes both the corresponding tree node and its sibling set, identical actions can receive different credit at different positions. The assigned safety credit thus depends jointly on the action, its history, its alternatives, and its observed consequences.

B THEORETICAL PROPERTIES OF TWO-SIDED SAFETY EVIDENCE

This section establishes the Bayesian interpretation, discriminative property, and robustness of the two-sided evidence defined in Eq. 2. Throughout, C denotes satisfaction of the target safety constraint, $\neg C$ denotes its violation, and μ is the behavior policy inducing the trajectory distribution.

Setup. Let s be an interaction history and let $a = (y_1, \ldots, y_L)$ be an action of length L. The safe reference trajectory c^+ and the corresponding violating reference trajectory c^- provide two privileged contexts for scoring the same action. They need not share a prefix with s. We use π_T and π_S solely as shorthand for the same scoring model π_ϕ under these two contexts:

$$\pi_T(a \mid s, c^+) \coloneqq \pi_\phi(a \mid s, c^+), \qquad \pi_S(a \mid s, c^-) \coloneqq \pi_\phi(a \mid s, c^-). \tag{7}$$

In particular, these are not independently trained scoring models. With $h_t^{\pm} = (s, c^{\pm}, y_{<t})$, autoregressive factorization gives

$$\begin{aligned} e(s,a) &= \sum_{t=1}^{L} \left[\log \pi_T(y_t \mid h_t^+) - \log \pi_S(y_t \mid h_t^-)\right] \\ &= \log \frac{\pi_\phi(a \mid s, c^+)}{\pi_\phi(a \mid s, c^-)}, \qquad \bar{e}(s,a) = \frac{e(s,a)}{L}. \end{aligned} \tag{8}$$

The notation $\bar{e}_\phi(s,a)$ in the main-text proposition refers to this same $\bar{e}(s,a)$.

Define the safety-success prior $\alpha_s \coloneqq p_\mu(C \mid s)$, with $0 < \alpha_s < 1$. The marginal action distribution satisfies

$$p_\mu(a \mid s) = \alpha_s p_\mu(a \mid s, C) + (1 - \alpha_s) p_\mu(a \mid s, \neg C). \tag{9}$$

We consider actions on the common support of the two outcome-conditioned distributions. The statistical approximation assumed below is

$$\begin{aligned} \left|\log \pi_\phi(a \mid s, c^+) - \log p_\mu(a \mid s, C)\right| &\le \epsilon_+, \\ \left|\log \pi_\phi(a \mid s, c^-) - \log p_\mu(a \mid s, \neg C)\right| &\le \epsilon_-. \end{aligned} \tag{10}$$

The presence of c^+ and c^- alone does not imply these inequalities; the results that invoke them hold when the stated approximation is satisfied.

Proposition B.1 (Properties of Two-Sided Safety Evidence). *Under the setup above, the two-sided evidence has the following properties.*

(i) Approximate Bayesian Evidence. *If Eq. 10 holds, then*

$$\left|\bar{e}(s,a) - \frac{\mathcal{B}(s,a)}{L}\right| \le \frac{\epsilon_+ + \epsilon_-}{L}, \tag{11}$$

where

$$\mathcal{B}(s,a) = \operatorname{logit} p_\mu(C \mid s, a) - \operatorname{logit} p_\mu(C \mid s).$$

(ii) Prior-Invariant Discrimination. *Holding the outcome-conditioned action distributions fixed, the one-sided contrast*

$$e_{\text{one}}(s,a) \coloneqq \log \frac{p_\mu(a \mid s, C)}{p_\mu(a \mid s)}$$

converges to zero as $\alpha_s \to 1$. In contrast, the ideal two-sided evidence

$$\mathcal{B}(s,a) = \log \frac{p_\mu(a \mid s, C)}{p_\mu(a \mid s, \neg C)} \tag{12}$$

does not explicitly depend on α_s. If all evaluated actions have the same length L, its expected class separation is

$$\begin{aligned} &\mathbb{E}_{a \sim p_\mu(\cdot \mid s, C)}\left[\frac{\mathcal{B}(s,a)}{L}\right] - \mathbb{E}_{a \sim p_\mu(\cdot \mid s, \neg C)}\left[\frac{\mathcal{B}(s,a)}{L}\right] \\ &\quad = \frac{1}{L}\Big[D_{\mathrm{KL}}\big(p_\mu(\cdot \mid s, C) \| p_\mu(\cdot \mid s, \neg C)\big) + D_{\mathrm{KL}}\big(p_\mu(\cdot \mid s, \neg C) \| p_\mu(\cdot \mid s, C)\big)\Big]. \end{aligned} \tag{13}$$

Under Eq. 10, the class separation computed with $\bar{e}(s,a)$ differs from Eq. 13 by at most $2(\epsilon_+ + \epsilon_-)/L$.

(iii) Robust Bounded Sensitivity. *Let*

$$\begin{aligned} \tilde{e}(s,a) &= \psi_{T,\delta}\big(\bar{e}(s,a)\big), \\ \psi_{T,\delta}(r) &= \operatorname{sign}(r) \tanh\left(\frac{(|r| - \delta)_+}{2T}\right) = \sigma_{T,\delta}(r), \end{aligned} \tag{14}$$

where $T > 0$, $\delta \ge 0$, and $(u)_+ = \max(u, 0)$. The last equality follows from the sigmoid definition of $\sigma_{T,\delta}$ in the main text. Then $|\tilde{e}(s,a)| \le 1$, $\psi_{T,\delta}(r) = 0$ for $|r| \le \delta$, and $\psi_{T,\delta}$ is $1/(2T)$-Lipschitz. If the token-level log-likelihood differences are perturbed by $\xi_1, \ldots, \xi_L$, the resulting score $\tilde{e}'(s,a)$ satisfies

$$|\tilde{e}'(s,a) - \tilde{e}(s,a)| \le \frac{1}{2TL} \sum_{t=1}^{L} |\xi_t|. \tag{15}$$

Proof. We establish the three properties in turn.

Proof of (i). Bayes' rule gives

$$\frac{p_\mu(C \mid s,a)}{p_\mu(\neg C \mid s,a)} = \frac{p_\mu(a \mid s,C)}{p_\mu(a \mid s,\neg C)} \frac{p_\mu(C \mid s)}{p_\mu(\neg C \mid s)}. \tag{16}$$

Taking logarithms and subtracting the prior log-odds yields

$$\mathcal{B}(s,a) = \operatorname{logit} p_\mu(C \mid s,a) - \operatorname{logit} p_\mu(C \mid s) = \log \frac{p_\mu(a \mid s,C)}{p_\mu(a \mid s,\neg C)}. \tag{17}$$

Define the two approximation residuals

$$\begin{aligned} r_+(s,a) &= \log \pi_\phi(a \mid s,c^+) - \log p_\mu(a \mid s,C), \\ r_-(s,a) &= \log \pi_\phi(a \mid s,c^-) - \log p_\mu(a \mid s,\neg C). \end{aligned} \tag{18}$$

Equations 8 and 17 imply

$$e(s,a) = \mathcal{B}(s,a) + r_+(s,a) - r_-(s,a).$$

Thus Eq. 10 and the triangle inequality give

$$|e(s,a) - \mathcal{B}(s,a)| \le |r_+(s,a)| + |r_-(s,a)| \le \epsilon_+ + \epsilon_-.$$

Dividing by L proves Eq. 11.

Proof of (ii). By Eq. 9, the one-sided contrast can be written as

$$\begin{aligned} e_{\mathrm{one}}(s,a) &= -\log\left[\alpha_s + (1-\alpha_s)\frac{p_\mu(a \mid s,\neg C)}{p_\mu(a \mid s,C)}\right] \\ &= -\log\left[\alpha_s + (1-\alpha_s)\exp\big(-\mathcal{B}(s,a)\big)\right]. \end{aligned} \tag{19}$$

For any action in the common support, this expression tends to zero as $\alpha_s \to 1$ when the conditional distributions are held fixed. By Eq. 12, $\mathcal{B}(s,a)$ contains no mixture weight α_s.

For the positive conditional distribution,

$$\mathbb{E}_{a\sim p_\mu(\cdot \mid s,C)}[\mathcal{B}(s,a)] = D_{\mathrm{KL}}\big(p_\mu(\cdot \mid s,C) \| p_\mu(\cdot \mid s,\neg C)\big). \tag{20}$$

For the negative conditional distribution,

$$\mathbb{E}_{a\sim p_\mu(\cdot \mid s,\neg C)}[\mathcal{B}(s,a)] = -D_{\mathrm{KL}}\big(p_\mu(\cdot \mid s,\neg C) \| p_\mu(\cdot \mid s,C)\big). \tag{21}$$

Subtracting these expectations and dividing by the fixed action length L gives Eq. 13. The sum of the two KL divergences is the Jeffreys divergence.

Finally, let

$$\Delta(s,a) = \bar{e}(s,a) - \frac{\mathcal{B}(s,a)}{L}.$$

By part (i), $|\Delta(s,a)| \le (\epsilon_+ + \epsilon_-)/L$. The difference between the class separation obtained from $\bar{e}$ and that obtained from $\mathcal{B}/L$ is therefore at most

$$\begin{aligned} &\mathbb{E}_{a\sim p_\mu(\cdot \mid s,C)}[|\Delta(s,a)|] + \mathbb{E}_{a\sim p_\mu(\cdot \mid s,\neg C)}[|\Delta(s,a)|] \\ &\quad \le \frac{2(\epsilon_+ + \epsilon_-)}{L}. \end{aligned} \tag{22}$$

Proof of (iii). The range of $\tanh$ implies $|\psi_{T,\delta}(r)| \le 1$. For $|r| \le \delta$, $(|r|-\delta)_+ = 0$, so $\psi_{T,\delta}(r) = 0$. Outside this dead zone, differentiation gives

$$\psi'_{T,\delta}(r) = \frac{1}{2T}\operatorname{sech}^2\left(\frac{|r|-\delta}{2T}\right), \qquad |r| > \delta. \tag{23}$$

The derivative is zero in the interior of the dead zone. Continuity at its boundaries and $\operatorname{sech}^2(x) \le 1$ therefore imply the global bound

$$\big|\psi_{T,\delta}(r') - \psi_{T,\delta}(r)\big| \le \frac{1}{2T}|r' - r|. \tag{24}$$

For $h_t^{\pm} = (s, c^{\pm}, y_{<t})$, define

$$z_t = \log \pi_T(y_t \mid h_t^{+}) - \log \pi_S(y_t \mid h_t^{-}). \tag{25}$$

Autoregressive factorization yields

$$e(s,a) = \sum_{t=1}^{L} z_t, \qquad \bar{e}(s,a) = \frac{1}{L}\sum_{t=1}^{L} z_t. \tag{26}$$

If $z'_t = z_t + \xi_t$, then

$$\left|\bar{e}'(s,a) - \bar{e}(s,a)\right| = \left|\frac{1}{L}\sum_{t=1}^{L}\xi_t\right| \le \frac{1}{L}\sum_{t=1}^{L}|\xi_t|.$$

Combining this inequality with Eq. 24 proves Eq. 15. In particular, if $|\xi_t| \le \eta$ for all t, then

$$|\tilde{e}'(s,a) - \tilde{e}(s,a)| \le \frac{\eta}{2T}.$$

If $|z_t| \le M$ for every token, then $|\bar{e}(s,a)| \le M$, whereas the unnormalized sum can have magnitude as large as LM. Thus token normalization removes this direct linear dependence on action length, and $\psi_{T,\delta}$ bounds the final score. □

Scope of the proposition. Proposition B.1 concerns statistical evidence under the trajectory distribution induced by μ. The Bayes factor measures how diagnostic an observed action is of C versus $\neg C$. Its approximation by reference-conditioned model scores remains conditional on Eq. 10; no matching between an arbitrary history s and the complete reference trajectories $c^{\pm}$ is required by the statement. The proposition alone does not establish a causal contribution of an individual action. Empirical assessments are provided next.

B.1 Empirical Validation of Two-Sided Safety Evidence

We examine the three properties in Proposition B.1. All privileged scores come from the same fixed model π_ϕ, evaluated with either c^+ or c^-. Figure 5 summarizes the results.

Bayesian-evidence calibration. For each evaluated state–action pair, we estimate the Bayesian log-odds update from held-out outcomes:

$$\widehat{\mathcal{B}}(s,a) = \operatorname{logit} \hat{p}_\mu(C \mid s,a) - \operatorname{logit} \hat{p}_\mu(C \mid s). \tag{27}$$

Because $\bar{e}(s,a)$ is token-normalized, we compare the sequence-level score $e(s,a) = L\bar{e}(s,a)$ with $\widehat{\mathcal{B}}(s,a)$. As shown in Figure 5(a), the two-sided estimator closely follows the identity line and achieves a Spearman correlation of $\rho = 0.99$. The one-sided estimator obtains $\rho = 0.94$ and exhibits greater dispersion, particularly for strongly positive evidence. These results support the calibration of the two-sided score against the estimated Bayesian update.

Discrimination under success-prior shift. We vary $\alpha_s = p_\mu(C \mid s)$ from 0.10 to 0.99 while holding the outcome-conditioned action distributions fixed. For an evidence score g, we measure expected class separation by

$$\mathcal{S}_{\alpha_s}(g) = \mathbb{E}_{a\sim p_\mu(\cdot\mid s,C)}[g(s,a)] - \mathbb{E}_{a\sim p_\mu(\cdot\mid s,\neg C)}[g(s,a)]. \tag{28}$$

For the fixed-length comparison in Figure 5(b), we use per-token scores: $e_{\text{one}}(s,a)/L$ and $\bar{e}(s,a)$. The one-sided separation decreases from approximately 0.053 at $\alpha_s = 0.10$ to nearly zero at $\alpha_s = 0.99$. The two-sided separation remains approximately 0.062 throughout this range and aligns with the independently computed per-token Jeffreys divergence. This comparison isolates the effect of changing the success-prior mixture weight.

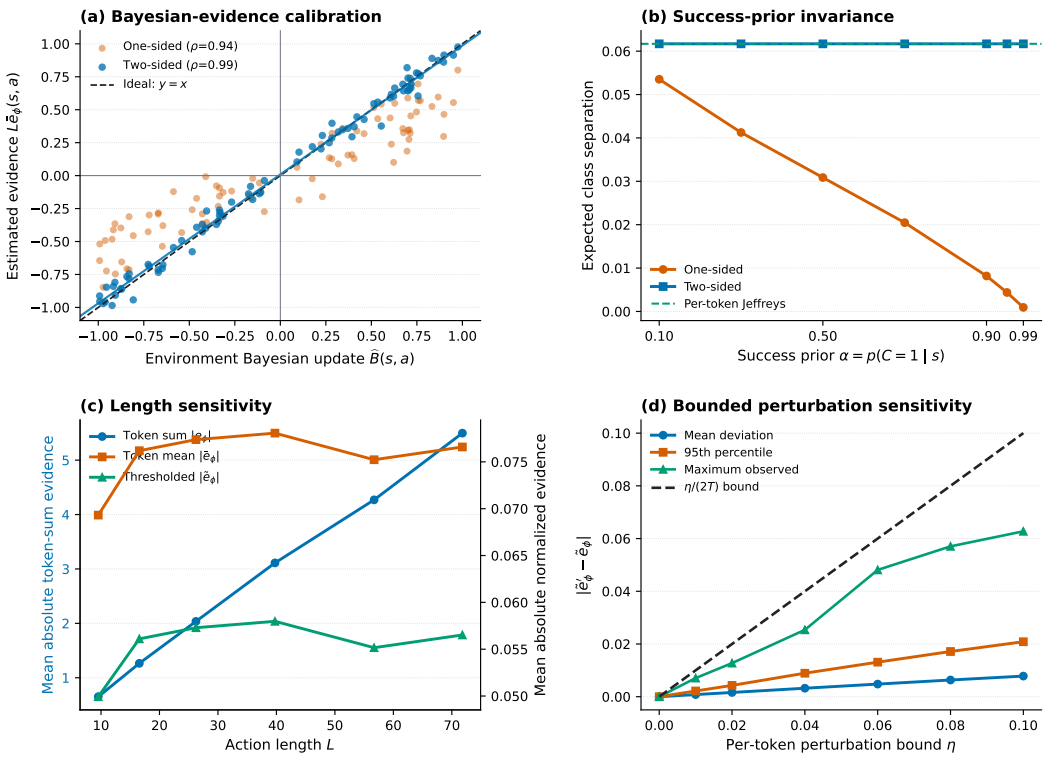

Figure 5: **Empirical validation of two-sided safety evidence. (a)** Calibration of the sequence-level score $e(s,a)$ against an environment-estimated Bayesian log-odds update. **(b)** Expected per-token class separation as the success prior changes while the outcome-conditioned action distributions remain fixed. **(c)** Action-length sensitivity of the token sum, normalized score, and thresholded score. **(d)** Sensitivity to bounded token-level perturbations; observed deviations remain below $\eta/(2T)$.

Length and perturbation robustness. Figure 5(c) compares $e(s,a)$, $\bar{e}(s,a)$, and $\tilde{e}(s,a)$ across six action-length groups. The magnitude of the token sum increases from approximately 0.65 to 5.5 as action length grows. The normalized score remains within approximately 0.069–0.078, and the thresholded score remains within approximately 0.050–0.057. The observed growth of the raw score is therefore substantially reduced by length normalization.

We also add perturbations satisfying $|\xi_t| \le \eta$ to the token-level log-likelihood differences. Equation 15 implies

$$|\tilde{e}'(s,a) - \tilde{e}(s,a)| \le \frac{\eta}{2T}.$$

With $T = 0.5$, Figure 5(d) shows that the mean, 95th-percentile, and maximum deviations all remain below this bound. At $\eta = 0.10$, these deviations are approximately 0.008, 0.021, and 0.063, respectively, compared with the upper bound of 0.10.

Taken together, the results support Bayesian-evidence calibration, discrimination under the specified prior shift, and robustness to action length and bounded token-level perturbations. They do not, by themselves, establish a causal effect for each scored action.

Outcome-conditioned likelihood approximation. Proposition B.1 assumes that π_ϕ under the two privileged contexts approximates the corresponding outcome-conditioned action distributions:

$$\begin{aligned} |\log \pi_\phi(a \mid s, c^+) - \log p_\mu(a \mid s, C)| &\le \epsilon_+, \\ |\log \pi_\phi(a \mid s, c^-) - \log p_\mu(a \mid s, \neg C)| &\le \epsilon_-. \end{aligned} \tag{29}$$

We assess this assumption on held-out interaction states using finite candidate action sets.

Table 4: Empirical approximation of the positive- and negative-outcome conditioned action distributions. Residuals are measured in log-probability space. Coverage reports the fraction of evaluated state–action pairs satisfying the indicated empirical error threshold.

Approximation	$\rho \uparrow$	Mean $\lvert r\rvert \downarrow$	Median $\lvert r\rvert \downarrow$	$\epsilon^{90} \downarrow$	$\epsilon^{95} \downarrow$
$\hat{\pi}_T$ vs. $\hat{p}_\mu(\cdot \mid s, C)$	0.97	0.10	0.07	0.21	0.27
$\hat{\pi}_S$ vs. $\hat{p}_\mu(\cdot \mid s, \neg C)$	0.96	0.12	0.08	0.24	0.31

For each state s, let $\mathcal{A}_s = \{a_1, \ldots, a_M\}$ be a set of distinct candidate actions sampled from the frozen behavior policy μ. For every a_j, we restore the environment at s, execute a_j, and sample R independent continuations from μ, following the counterfactual protocol in Appendix D.2. This yields

$$\hat{q}_j = \hat{p}_\mu(C \mid s, a_j) = \frac{1}{R} \sum_{r=1}^{R} S(\tau_r^{s,a_j,\mu}). \tag{30}$$

By Bayes' rule, the estimated outcome-conditioned distributions, restricted and renormalized to $\mathcal{A}_s$, are

$$\begin{aligned} \hat{p}_{+,\mathcal{A}_s}(a_j \mid s) &= \frac{\mu(a_j \mid s)\hat{q}_j}{\sum_{\ell=1}^{M} \mu(a_\ell \mid s)\hat{q}_\ell}, \\ \hat{p}_{-,\mathcal{A}_s}(a_j \mid s) &= \frac{\mu(a_j \mid s)(1-\hat{q}_j)}{\sum_{\ell=1}^{M} \mu(a_\ell \mid s)(1-\hat{q}_\ell)}. \end{aligned} \tag{31}$$

For each state, we hold its reference pair $c^{\pm}$ fixed across the candidates and normalize the corresponding model likelihoods over the same set:

$$\begin{aligned} \hat{\pi}_{T,\mathcal{A}_s}(a_j \mid s, c^+) &= \frac{\pi_\phi(a_j \mid s, c^+)}{\sum_{\ell=1}^{M} \pi_\phi(a_\ell \mid s, c^+)}, \\ \hat{\pi}_{S,\mathcal{A}_s}(a_j \mid s, c^-) &= \frac{\pi_\phi(a_j \mid s, c^-)}{\sum_{\ell=1}^{M} \pi_\phi(a_\ell \mid s, c^-)}. \end{aligned} \tag{32}$$

Log residuals are computed for candidates with positive estimated masses:

$$\begin{aligned} r_{+,\mathcal{A}_s}(s, a_j) &= \log \hat{\pi}_{T,\mathcal{A}_s}(a_j \mid s, c^+) - \log \hat{p}_{+,\mathcal{A}_s}(a_j \mid s), \\ r_{-,\mathcal{A}_s}(s, a_j) &= \log \hat{\pi}_{S,\mathcal{A}_s}(a_j \mid s, c^-) - \log \hat{p}_{-,\mathcal{A}_s}(a_j \mid s). \end{aligned} \tag{33}$$

For each direction, we report rank agreement, the mean and median absolute residuals, and the empirical 95th-percentile absolute residual

$$\epsilon^{95}_{\pm,\mathcal{A}} = Q_{0.95}(|r_{\pm,\mathcal{A}_s}(s, a)|)\,. \tag{34}$$

Table 4 shows that the privileged scores track the corresponding outcome-conditioned distributions on the evaluated candidate sets. For the safety-conditioned distribution, the positive-context scores achieve a Spearman correlation of 0.97, a mean absolute log residual of 0.10, and a 95th-percentile residual of 0.27. The corresponding negative-context values are 0.96, 0.12, and 0.31. The larger residual for the negative condition is consistent with greater heterogeneity among violating continuations.

These candidate-set residuals assess agreement after context-specific normalization. Their 95th percentiles are empirical diagnostics, not uniform bounds on the original, unnormalized log likelihoods in Eq. 10. Together with the Bayesian calibration in Figure 5, they provide evidence for the approximation in the evaluated setting, while leaving its validity for arbitrary histories, reference pairs, and actions as an explicit assumption.

C Detailed Experimental Setup

C.1 Benchmarks

The main paper reports aggregate results on three benchmarks. Here we explain the distinct safety property measured by each benchmark and how the three evaluations complement one another.

AgentHarm. **AgentHarm** (Andriushchenko et al., 2025) tests whether an agent rejects explicitly harmful requests while preserving its ability to complete benign multi-step tasks. We use its harmful and benign splits together so that improvements caused by indiscriminate refusal are separated from genuine movement of the safety utility frontier.

Agent-SafetyBench. **Agent-SafetyBench** (Zhang et al., 2024) covers multiple interactive environments, risk categories, and agent failure modes. Unlike AgentHarm, its risks may emerge only after interaction or tool use begins. We report the aggregate safety score together with category-level results to distinguish broad improvements from gains limited to a few risk types.

SafeAgentBench. **SafeAgentBench** (Yin et al., 2024) evaluates planning and execution in environments where the requested objective and safety constraints must both be satisfied throughout a trajectory. We evaluate its detailed and abstract task settings with the official execution-oriented metrics. For the long-horizon study, we additionally select tasks in which several turns separate a safety-relevant decision from its consequence. This subset tests whether credit remains reliable as delay increases and whether the signal captures dependencies created by action order.

Complementarity of the benchmarks. AgentHarm isolates harmful-request handling and benign-task retention; Agent-SafetyBench broadens the evaluation to heterogeneous interaction risks; and SafeAgentBench stresses constrained execution with delayed effects. Their combination evaluates both final behavior and the sequential credit-assignment property targeted by HFDTree. These three datasets natively provide paired safe and unsafe trajectories.

C.2 Baselines

Our baseline suite varies the source and granularity of safety supervision. It includes the unaligned backbone, outcome-supervised alignment, privileged-information methods, turn-level estimators, and agent-specific safety methods. This organization tests whether the reported gains follow from the proposed credit construction rather than from post-training alone, access to privileged context, or additional safety data.

Base Model. The **Base Model** is the unaligned student backbone and provides the reference for measuring the effect of safety post-training.

Safe RLHF-v. **Safe RLHF-v** (Ji et al., 2026a) represents outcome-supervised safety alignment. It optimizes trajectory-level safety feedback without explicitly decomposing the outcome into decision-specific temporal credit. The comparison therefore measures what turn-level attribution adds beyond terminal feedback.

Privileged-information credit baselines. **OPSD** (Zhao et al., 2026a) and **PBSD** (Tian et al., 2026) derive fine-grained supervision from teacher-side or otherwise privileged information. They control for the benefit of richer context. HFDTree uses the same general resource for a different purpose: estimating temporally structured safety contributions for individual decisions.

Turn-level credit baselines. **GraphGPO** (Cheng et al., 2026), and **AT^{2}PO** (Zong et al., 2026) assign credit more finely than a trajectory-level reward. They provide the closest controls for identifying which intermediate decisions produce later outcomes. HFDTree additionally conditions safety credit on both the preceding interaction and future consequences, so the same action may receive different credit across states and consequence delays.

Agent safety alignment baselines. **AgentAlign** (Zhang et al., 2025) and **SafeBranch** (Lee et al., 2026) directly target LLM-agent safety. They control for the effect of training on safety-oriented interactions, which separates the proposed credit mechanism from generic benefits of agent-specific safety post-training.

Controlled comparison. Directly comparable methods share the student backbone, task split, rollout budget, and terminal safety feedback. When a method requires a teacher or privileged context,

we also match the teacher configuration and available supervision as closely as its formulation allows. No baseline receives a larger interaction budget, a stronger backbone, or extra safety labels unless required by its original procedure. The remaining differences therefore primarily concern supervision and credit assignment.

Implementation and reproducibility details. Unless otherwise specified, all models are optimized with AdamW using a learning rate of 1×10^{-5}, a weight decay of 0.01, and $1{,}000$ optimization steps. We use a rollout group size of $G = 8$ and set the HFDTree hyperparameters to $\gamma = 0.9$, $T = 0.5$, $\delta = 0.05$, and $\epsilon = 0.05$, selected on a held-out validation split and fixed for all test evaluations. Training is conducted on $8\times$ NVIDIA A100 80GB GPUs. For the diagnostic experiments, we use 256 held out state action pairs for counterfactual credit evaluation, 128 paired-prefix interventions for context sensitivity, 200 action order comparisons, and 256 state action pairs for the outcome-conditioned likelihood approximation. Each counterfactual action is evaluated with $R = 32$ independently sampled continuations. The delayed-consequence evaluation contains 10 cases with delays ranging from 2 to 16 turns. Main benchmark results are averaged over three independent runs, and all methods within each comparison use the same decoding configuration, rollout budget, interaction budget, and diagnostic instances.

C.3 LLM Backbones

Default student and teacher. The default student and teacher is **Qwen3-4B**. The frozen teacher supplies the positive and negative behavioral evidence used to derive turn-level credit; optimization updates only the student. This separation transfers the teacher's stronger safety assessment without modifying the teacher itself.

Self-distillation control. The **self-distillation** control uses the same backbone for teacher and student, while only the teacher receives the privileged context required by the supervision procedure. Improvements in this setting isolate the credit formulation from any capacity advantage of a larger teacher.

Controlled model settings. Within each comparison group, methods share the student backbone, teacher model, decoding configuration, rollout-group size, and interaction budget whenever the baseline permits. These controls prevent sampling and budget differences from being conflated with the alignment objective.

C.4 Evaluation Metrics

The reported metrics address three distinct questions: whether the trained agent is safe and useful, whether turn-level credit matches counterfactual safety contribution, and whether the estimator responds to safety-relevant changes in action order.

C.4.1 Benchmark-Level Safety and Utility Metrics

AgentHarm. For AgentHarm, we report **Safety Violation Rate (SVR)**, **Over-Refusal Rate (ORR)**, and **Safe Completion Rate (SCR)**. SVR measures how frequently an agent produces or executes behavior that violates the benchmark's safety criteria, whereas ORR measures unnecessary refusals on benign or safely executable requests. SCR counts episodes that satisfy the task-success criterion without incurring any safety violation. Reporting these metrics jointly distinguishes genuine safety improvements from reductions in harmful behavior achieved through indiscriminate refusal.

Agent-SafetyBench. For Agent-SafetyBench, we report SVR, ORR, and SCR over the complete evaluation set and additionally examine their category-level breakdowns. SVR captures unsafe behavior across the benchmark's risk categories, ORR reflects excessive conservatism on permissible tasks, and SCR measures successful and violation-free task completion. The category-level results indicate whether an improvement generalizes across heterogeneous safety risks or is concentrated in a limited subset of scenarios.

SafeAgentBench. For both the detailed and abstract tasks in SafeAgentBench, we report SVR, ORR, and SCR under the benchmark's trajectory-level evaluation protocol. For long-horizon tasks, safety is evaluated over the complete execution trajectory rather than solely from the final outcome. An episode contributes to SVR if any executed action violates a safety constraint, while it contributes to SCR only if the task is completed without any such violation. Consequently, a trajectory that reaches a nominally successful final state through an unsafe intermediate action is excluded from SCR and counted as a safety violation. ORR separately captures safe tasks that are unnecessarily rejected or terminated. Lower SVR and ORR indicate better behavior, whereas higher SCR indicates a stronger safety–utility trade-off. "'

C.4.2 Turn-Level Credit Evaluation

We evaluate attribution directly by intervening on individual turns and measuring the resulting change in trajectory safety.

For a trajectory

$$\tau = (s_1, a_1, \ldots, s_T, a_T),$$

we replace the action at turn t with an alternative a'_t, preserve the prefix $s_1, a_1, \ldots, s_t$, and re-execute the environment from that decision. This produces a counterfactual trajectory with an independently observed safety outcome.

Let $\hat{c}_t$ denote the credit assigned by the evaluated method to turn t, and let

$$\Delta S_t = S(\tau) - S(\tau^{(t)})$$

denote the change in trajectory safety between the original trajectory and the corresponding counterfactual trajectory $\tau^{(t)}$. Agreement between $\hat{c}_t$ and ΔS_t measures whether the estimator recovers the decision's observed contribution to future safety.

Spearman correlation. We report the **Spearman rank correlation**

$$\rho_{\text{credit}} = \text{Spearman}\left(\{\hat{c}_t\}, \{\Delta S_t\}\right),$$

between predicted credit and counterfactual contribution. Larger values mean that the estimator ranks turns more consistently with their observed safety effects.

Top-k Recall. **Top-k Recall** is the fraction of ground-truth critical decisions among the k turns with the largest predicted credit magnitude. It measures retrieval quality when only a few turns in a long trajectory determine safety.

C.4.3 Action Order Sensitivity

Safety can depend on temporal organization rather than merely on which actions appear. We therefore test whether an estimator distinguishes different orderings of the same action set.

Each **order-perturbed trajectory pair** preserves the task and action set but changes the execution order of safety-relevant actions. The pair controls for action identity and isolates sensitivity to temporal structure.

Let τ^+ denote the safer ordering and τ^- denote its unsafe or less-safe permutation. We report **Order Preference Accuracy**:

$$\text{OPA} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[R(\tau_i^+) > R(\tau_i^-)\right],$$

where $R(\cdot)$ is the trajectory reward or preference score produced by the evaluated method. A larger value indicates that the learned signal prefers the safer temporal organization rather than responding only to action occurrence.

C.4.4 Summary of Evaluation Dimensions

In summary, benchmark metrics measure the safety and utility of the trained policy; counterfactual metrics test decision-level attribution; and order preference tests whether credit reflects temporal dependencies. The latter two diagnostics directly evaluate the properties that motivate HFDTree rather than relying only on end-to-end performance.

Table 5: Tree structure statistics under the training rollout configuration. R_{branch} measures non root branching, R_{center} reports the fraction of training nodes receiving sibling centered credit, B_{eff} is the average sibling set size among branching nodes, and D_{branch} is the deepest branching position normalized by trajectory horizon.

Benchmark	$R_{\text{branch}} \uparrow$	$R_{\text{center}} \uparrow$	$B_{\text{eff}} \uparrow$	$D_{\text{branch}} \uparrow$
AgentHarm	61.8%	73.5%	3.7	0.68
Agent-SafetyBench	58.4%	69.2%	3.5	0.64
SafeAgentBench	64.1%	76.8%	3.9	0.72
Average	61.4%	73.2%	3.7	0.68

D Extended Experiments

The experiments below complement the end-to-end results in Section 3.2. They test whether safety training preserves general agent capability and whether the learned turn-level credit agrees with independent counterfactual interventions.

D.1 Tree Structure and Non-degeneration

Motivation. HFDTree relies on comparing alternative continuations under a common interaction prefix. A tree constructed by post-hoc matching of independently generated free-form trajectories could degenerate after the first action, because different actions induce different subsequent textual histories. Our rollout procedure instead explicitly branches from checkpointed interaction prefixes. We therefore examine whether the resulting trees contain substantial branching beyond the root under the actual training configuration.

Tree-structure metrics. For each rollout tree T_x, let $\mathrm{ch}(p)$ denote the children of node p and $d(p)$ its depth. We report four complementary statistics.

The non-root branching rate is

$$R_{\text{branch}} = \frac{|\{p : d(p) > 0,\ |\mathrm{ch}(p)| > 1\}|}{|\{p : d(p) > 0,\ |\mathrm{ch}(p)| \geq 1\}|},$$

which measures how frequently intermediate decision nodes contain explicit alternatives.

More directly, we define the sibling-centering coverage

$$R_{\text{center}} = \frac{|\{n : |\mathrm{ch}(p(n))| > 1\}|}{|\mathcal{N}| - 1},$$

the fraction of non-root training nodes for which Eq. (4) uses the sibling-centered credit $B_n - \bar{B}_{p(n)}$ rather than the singleton fallback B_n.

We additionally report the average sibling-set size

$$B_{\text{eff}} = \mathbb{E}[|\mathrm{ch}(p)| \mid |\mathrm{ch}(p)| > 1,\ d(p) > 0],$$

and the effective branching depth

$$D_{\text{branch}} = \max_{p:\,|\mathrm{ch}(p)|>1} d(p),$$

normalized by the trajectory horizon when trajectories have different lengths.

Results. Table 5 shows that the rollout trees retain substantial branching beyond the root. Across the three benchmarks, 58.4–64.1% of non-root internal nodes contain multiple continuations, and 69.2–76.8% of training nodes therefore receive sibling-centered credit rather than the singleton fallback in Eq. (4). Branching nodes contain 3.5–3.9 children on average and remain present until approximately 64–72% of the trajectory horizon. These statistics indicate that HFDTree does not degenerate into independent chains after the root; sibling comparison remains active across intermediate decision states.

D.2 General Agent Capability Retention

Motivation. Safety training can reduce violations by making an agent overly conservative, which may also weaken tool use and benign-task completion. We measure this possible *alignment tax* by evaluating the normally trained checkpoint from each method on both safety tasks and independent capability benchmarks. Unlike a matched operating-point analysis, this experiment does not tune checkpoints to reach a predetermined safety rate.

Evaluation protocol. We evaluate Qwen3-4B and LLaMA-3.1-8B-Instruct. Within each backbone, the unaligned model and every safety-aligned variant share the architecture, agent scaffold, system prompt, tool specifications, context limit, maximum number of turns, and decoding configuration; only post-training differs. General-agent benchmark examples are excluded from safety training, checkpoint selection, and hyperparameter tuning. All checkpoints use the same benchmark implementation and evaluator. We report means and standard deviations over three independent runs for stochastic environments and use an identical decoding setup for deterministic function-calling tasks.

D.2.1 General Agent Capability Benchmarks

Berkeley Function Calling Leaderboard (BFCL v3). BFCL v3 (Patil et al., 2025) measures whether a model selects and invokes external functions correctly. Its single-turn split tests local tool choice, whereas its multi-turn and multi-step split also requires context tracking and sequenced function calls. We use the official evaluator and report **BFCL-ST**↑ and **BFCL-MT**↑ separately.

τ-bench. τ-bench (Yao et al., 2025) evaluates policy-compliant API use during multi-turn exchanges with a simulated user. We run the Retail and Airline domains with the official database-state evaluator and report **pass**^{1}↑, averaged across domains. This task complements BFCL by testing sustained tool use across a longer interaction.

AgentBench. AgentBench (Liu et al., 2024) covers several interactive environments that require multi-turn reasoning, planning, and decision making. We use the official environment-specific evaluators and report the aggregate **AgentBench score**↑, which broadens the capability assessment beyond function calling.

D.2.2 Metrics

Safety metrics. The safety side of this comparison uses SafeAgentBench. We report safety violation rate (**SVR**↓), over-refusal rate (**ORR**↓), and safe completion rate (**SCR**↑). These metrics respectively measure unsafe execution, unnecessary conservatism, and successful completion under all required safety constraints.

Capability change. For each general-agent metric b, the change from the corresponding unaligned backbone is

$$\Delta_{\text{cap}}^{(b)}(m) = S_b(m) - S_b(\text{Base}), \tag{35}$$

where $S_b(m)$ denotes the score of alignment method m on benchmark b. Values near zero indicate capability retention, negative values indicate an alignment tax, and positive values indicate improvement after alignment.

For a compact summary, we average the changes across the four reported capability metrics:

$$\overline{\Delta}_{\text{cap}}(m) = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \Delta_{\text{cap}}^{(b)}(m), \qquad \mathcal{B} = \{\text{BFCL-ST}, \text{BFCL-MT}, \tau\text{-bench}, \text{AgentBench}\}. \tag{36}$$

All four scores use a 0–100 scale, so $\overline{\Delta}_{\text{cap}}$ is measured in percentage points. The average summarizes retention, while the individual scores remain the primary evidence.

D.2.3 Results

Table 6 reports safety and general capability for the same checkpoints. The Base Model is the backbone before safety training; Safe RLHF-v supplies an outcome-supervised baseline; and AT^{2}PO

Table 6: **Safety improvement and general agent capability retention across two backbones.** Safety metrics are evaluated on SafeAgentBench. BFCL-ST, BFCL-MT, τ-bench, and AgentBench evaluate general agent capabilities outside the safety-alignment benchmarks. $\overline{\Delta}_{\mathrm{cap}}$ denotes the average percentage point change over the four capability metrics relative to the corresponding unaligned backbone. Best aligned model results within each backbone are shown in bold.

		SafeAgentBench			General Agent Capability				
Backbone	**Method**	**SVR↓**	**ORR↓**	**SCR↑**	**BFCL-ST↑**	**BFCL-MT↑**	**τ-bench↑**	**AgentBench↑**	**$\overline{\Delta}_{\mathrm{cap}}$**
Qwen3-4B	Base Model	42.1±1.2	8.9±0.6	11.8±0.7	57.6	46.3	51.2	39.7	–
	Safe RLHF-v	28.1±1.0	8.6±0.5	25.0±0.8	53.1	42.4	48.0	37.6	−3.4
	AT^{2}PO	17.5±0.8	6.8±0.4	36.2±0.7	43.8	35.9	45.2	38.1	−8.0
	HFDTree	**13.8±0.7**	**6.1±0.3**	**41.9±0.7**	**57.3**	**47.6**	**50.9**	**39.5**	+0.1
LLaMA-3.1-8B	Base Model	43.6±1.2	9.4±0.6	10.7±0.7	84.2	9.6	16.1	36.5	–
	Safe RLHF-v	29.7±1.0	9.1±0.5	22.7±0.8	80.0	8.8	15.1	34.6	−2.0
	AT^{2}PO	18.9±0.8	7.2±0.4	33.6±0.7	81.4	9.0	15.3	35.2	−1.4
	HFDTree	**15.0±0.7**	**6.4±0.3**	**38.7±0.7**	**83.9**	**9.8**	**16.0**	**36.3**	−0.1

represent fine-grained or tree-structured credit assignment. No aligned checkpoint is adapted further on the capability benchmarks.

Safety capability trade-off. This analysis compares each normally trained checkpoint with its unaligned backbone rather than selecting an artificial matched-safety operating point. For method m, the reduction in safety violations is

$$\Delta_{\mathrm{safety}}(m) = \mathrm{SVR}(\mathrm{Base}) - \mathrm{SVR}(m), \tag{37}$$

while Eq. 35 measures the corresponding capability change. A useful alignment method should make Δ_{safety} large and positive while keeping $\Delta^{(b)}_{\mathrm{cap}}$ close to zero.

Discussion. HFDTree reduces SVR by 28.3 percentage points for Qwen3-4B and 28.6 points for LLaMA-3.1-8B while increasing SCR by 30.1 and 28.0 points, respectively. Its four capability scores remain within 1.3 points of the corresponding base model, producing mean changes of $+0.1$ and -0.1 points. AT^{2}PO, by contrast, has mean changes of -80. These results show that stronger safety need not require a broad reduction in tool use or general agent capability.

Reproducibility details. BFCL v3 uses its official function-calling format and evaluator. Across τ-bench checkpoints, we fix the user simulator, tool definitions, policy documents, and interaction budget for both Retail and Airline. AgentBench runs share environment configurations, prompt templates, action parsers, and interaction budgets. Every capability number comes from our reruns under matched settings rather than from cross-leaderboard comparisons.

D.3 Counterfactual Evaluation of Turn-Level Credit

Evaluation setup. All credit estimators are evaluated on the same held-out pool of interaction prefixes and actions. We freeze the student continuation policy and every scoring model, and we keep evaluation templates separate from training and score calibration. Each method receives identical instances and rollout information. The protocol therefore compares attribution quality under fixed agent behavior, independently of end-to-end policy performance.

Counterfactual safety advantage. Let $S(\tau) \in \{0, 1\}$ indicate whether trajectory τ satisfies every safety constraint throughout execution. Under a fixed continuation policy μ, the action value is

$$Q^{\mu}_{\mathrm{safe}}(s, a) = \mathbb{E}[S(\tau) \mid s, a, \mu]\,. \tag{38}$$

We estimate it by restoring the environment at prefix s, executing a, and sampling R continuations from μ:

$$\widehat{Q}^{\mu}_{\mathrm{safe}}(s, a) = \frac{1}{R} \sum_{r=1}^{R} S(\tau_r^{s,a,\mu}). \tag{39}$$

Given a reference distribution $q(a' \mid s)$ over valid alternatives, the estimated counterfactual safety advantage is

$$A_{\mathrm{CF}}(s,a) = \widehat{Q}^{\mu}_{\mathrm{safe}}(s,a) - \sum_{a'} q(a' \mid s)\widehat{Q}^{\mu}_{\mathrm{safe}}(s,a'). \tag{40}$$

We hold the candidate construction, reference distribution, continuation policy, and execution budget fixed across methods. Reference continuations are sampled independently of the rollouts used to construct HFDTree and the other credit estimates. Accordingly, A_{CF} is an action's contribution relative to the chosen alternatives and policy, not an intrinsic context-free value.

Critical-turn capture. For evaluated turn i, let $c_{m,i}$ be the raw credit from method m. We use Δ_n before optimization-time clipping for HFDTree and $\tilde{e}(s,a)$ for its local variant. Let $\mathcal{I}_{m,b}$ be the top b fraction of turns ranked by $|c_{m,i}|$. Their contribution capture is

$$\mathrm{Capture}_m(b) = \frac{\sum_{i \in \mathcal{I}_{m,b}} |A_{\mathrm{CF},i}|}{\sum_i |A_{\mathrm{CF},i}|}. \tag{41}$$

An oracle ranks turns by $|A_{\mathrm{CF},i}|$, and every method is evaluated on the same turn pool.

Paired prefix interventions. Each intervention pairs two prefixes, s_0 and s_1, under which the same action a can be executed. The pair changes a safety-relevant property of the history but preserves the evaluated action, continuation policy, and reference distribution over alternatives. For example, an identical file-sharing action can follow a history where sensitive content was either redacted or retained. Neutral pairs change only safety-irrelevant context. We fix pair orientation and control membership before examining model predictions.

For each pair, we compute

$$\delta A_{\mathrm{CF}} = A_{\mathrm{CF}}(s_1,a) - A_{\mathrm{CF}}(s_0,a), \qquad \delta \hat{c}_m = \hat{c}_m(s_1,a) - \hat{c}_m(s_0,a). \tag{42}$$

Spearman correlation evaluates the ranking of these shifts, while sign agreement evaluates their direction. For safety-relevant pairs, we also report prefix-shift MAE:

$$\mathrm{MAE}_{\mathrm{prefix},m} = \frac{1}{N_{\mathrm{pair}}} \sum_i \left| \delta \hat{c}_{m,i} - \delta A_{\mathrm{CF},i} \right|. \tag{43}$$

Neutral controls are reported separately so that insensitivity to irrelevant changes cannot inflate agreement on safety-relevant pairs.

Score scaling for absolute errors. We additionally compare against TRACE (Hong et al., 2026), a state-of-the-art safety detection method.

Raw credit and counterfactual advantage may use different scales. For MAE, we therefore transform credit as $\hat{c}_m = \alpha_m c_m$, where $\alpha_m > 0$ is fitted on a disjoint calibration set by minimizing absolute error against A_{CF} and then frozen. The same coefficient applies to both prefixes in every pair and to all delayed-consequence cases; no condition-specific or test-set calibration is allowed. This positive, zero-intercept transformation preserves ordering and sign while placing absolute errors on the safety-advantage scale. Capture curves continue to use raw magnitudes.

Delayed-consequence attribution. The delayed-consequence set contains ten cases in which the relevant outcome occurs 2–16 turns after the evaluated action. For each case, we compute the earlier action's advantage with Eq. 40 and compare it with each method's credit. Attribution error is

$$E_{m,i} = \left| \hat{c}_{m,i} - A_{\mathrm{CF},i} \right|, \qquad \mathrm{MAE}_{\mathrm{delay},m} = \frac{1}{10} \sum_{i=1}^{10} E_{m,i}. \tag{44}$$

Cases are selected for their delayed safety dependency before method errors are compared. This analysis evaluates attribution on those cases; it does not treat cross-case differences as the isolated empirical evidence effect of delay. The controlled delay sweep provides that complementary test.

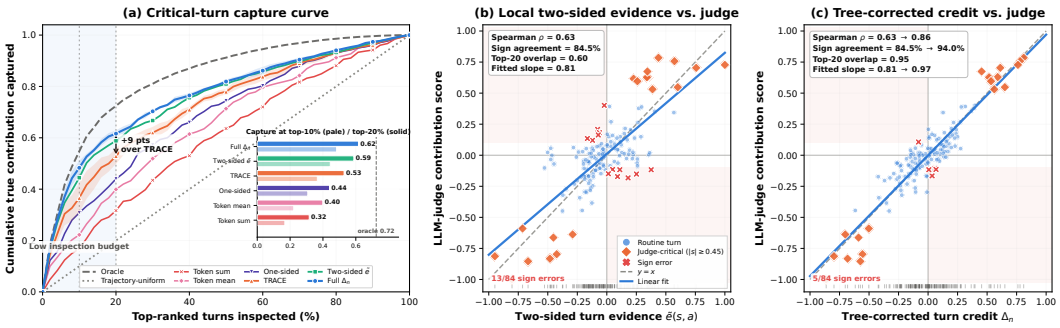

Figure 6: Validation of turn-level safety credit. (a) Cumulative absolute counterfactual contribution captured by ranking turns by predicted credit magnitude. (b,c) Agreement of local evidence $\tilde{e}(s, a)$ and tree-corrected credit Δ_n with signed contribution scores from an independent LLM judge.

Uncertainty and interpretation. Repeated continuations reduce uncertainty in the environment-based reference. Confidence intervals resample independent tasks while keeping turns and intervention pairs from the same task together. When finite-continuation noise is non-negligible, continuation outcomes are additionally resampled within each task. Correlation supports ranking agreement, while absolute-error claims depend on the scaling protocol. Neither result implies that unscaled credit is a calibrated empirical evidence effect.

D.4 Validity of Turn-Level Credit

HFDTree prioritizes turns with substantial effects on downstream safety. Figure 6a ranks turns by predicted credit magnitude and measures the fraction of absolute counterfactual contribution recovered. At a 20% inspection budget, Δ_n captures 9 percentage points more contribution than the strongest prior method. Two-sided evidence accounts for most of the gain, improving capture by 16 points over one-sided evidence, and tree correction adds another 2 points. The success- and failure-conditioned contrast in Eq. 2 therefore helps identify actions whose intervention changes future safety.

Tree correction improves the ranking and direction of turn-level credit. An independent judge provides a complementary signed-attribution comparison. Relative to local two-sided evidence, tree correction raises Spearman correlation by 0.23 and sign agreement by 9.5 percentage points. It therefore improves both the ordering of contributions and the predicted direction of their safety effect. Combined with the capture result, this finding separates the roles of the two components: two-sided evidence locates consequential turns, and tree correction refines their context-dependent credit.

D.5 Ablation Protocol and Evaluation Metrics

Component configurations. Five configurations separate evidence construction from credit assignment. *One-sided + HFDTree* retains propagation and sibling centering but uses one-sided evidence. *Two-sided local only* removes both tree operations. *Two-sided + future* propagates evidence without centering, whereas *Two-sided + centering* centers local evidence without propagation. *Full HFDTree* first propagates two-sided evidence and then centers the resulting branch values. Comparisons with the one-sided, centering-only, and future-only variants isolate evidence construction, propagation, and centering, respectively. is an external baseline, not a controlled component ablation.

Evaluation organization. All variants use a shared diagnostic pool and the same reference judgments. These diagnostics measure attribution quality. A separate SafeAgentBench evaluation measures the safety and utility of Qwen3-4B policies trained with each variant, distinguishing accurate credit from the behavior produced by training on that credit.

Counterfactual correlation (CF ρ). CF ρ compares predicted turn credit with a reference contribution derived from counterfactual environment outcomes. A larger correlation indicates that the estimator orders turns more consistently with their measured safety effects.

Delayed-consequence accuracy (Delayed Acc.). Delayed Acc. is the percentage of delayed-consequence comparisons for which the credit prediction matches the reference judgment. It tests whether an estimator attributes downstream effects that may be absent from local evidence.

Action-order preference accuracy (Order Acc.). Order Acc. is the percentage of safety-relevant ordering comparisons for which the credit-based preference matches the reference. The metric tests whether an action's predicted contribution changes appropriately with its preceding context.

Trained-policy safety. Trained policies are evaluated with SVR, ORR, and SCR. Lower SVR means fewer safety violations, lower ORR means fewer unnecessary refusals, and higher SCR means more successful completions without violations. Reporting the three together prevents excessive refusal from being mistaken for improved safety.

Reporting conventions. Accuracy and rate metrics are percentages, CF ρ is unitless, and differences between percentages are stated in percentage points. We mark unavailable or inapplicable entries with "–".