RISA: RISK-INFORMED SAFETY ACTIVATION FOR LONG-HORIZON TOOL USING AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Tool-using language-model agents increasingly operate over long interaction horizons, where unsafe outcomes may emerge only after a sequence of individually plausible actions. Existing safeguards often focus on the current action or compress trajectory history into a single risk decision, providing limited control over when and how strongly to intervene. We introduce **RISA** (**R**isk-**I**nformed **S**afety **A**ctivation), a framework for long-horizon agent safety that combines multi-horizon future-risk prediction with adaptive representation steering. RISA models the distribution over the first future safety violation, derives risk estimates across multiple horizons, and uses these predictions to train an intervention policy that adaptively controls the strength of a safety-relevant internal representation direction. Complete trajectories, consisting of the full sequence of agent inputs, actions, and tool or environment observations, provide privileged supervision during training, while inference depends only on the observed trajectory prefix. Across multiple long-horizon agent-safety benchmarks and model families, RISA improves unsafe completion and safe completion while preserving task utility and limiting over-refusal compared with step-level, trajectory-aware, fixed-steering, and threshold-based safeguards. Additional analyses show that the learned safety direction captures behaviorally relevant structure, multi-horizon conditioning supports more selective intervention, and RISA shifts risky trajectory states toward lower predicted future risk. These results suggest that separating future-risk forecasting from adaptive internal control provides an effective approach to proactive safety intervention in long-horizon agents.

1 INTRODUCTION

Large language models (LLMs) are increasingly deployed as tool-using agents that interact with external environments over extended trajectories. Such agents can read and modify files, query databases, execute code, invoke APIs, and make decisions whose effects persist beyond a single response (Mou et al., 2026; Li et al., 2026b; Lin et al., 2026b). These capabilities create safety challenges that differ from those of standalone text generation (Campbell et al., 2026; Liu et al., 2026; Asif & Amiri, 2026; Karamat et al., 2026). In long-horizon settings (Dong et al., 2026), unsafe outcomes may emerge gradually: actions that appear permissible in isolation can collectively move the environment toward a state in which a later action violates a safety constraint. Detecting only the final unsafe action can therefore leave little opportunity to redirect the trajectory before the violation occurs.

Many agent safeguards focus primarily on the action currently under consideration (Xiang et al., 2025; Chen et al., 2025). They inspect the current interaction context or proposed tool call and determine whether the action should be allowed, blocked, or refused. Such mechanisms are well suited to cases in which unsafe intent or behavior is already visible in the pending action, but they are less informative when risk develops across multiple otherwise plausible steps. For example, reading a file, retrieving account information, and preparing data for transfer may each be individually permissible while collectively moving the agent toward an unsafe disclosure. In these cases, safety depends not only on the current action but also on how the trajectory is evolving.

Recent work has therefore incorporated longer interaction histories into agent-safety decisions. ToolSafe (Mou et al., 2026) evaluates tool use in interaction context, TRACE (Hong et al., 2026) aggregates safety-relevant evidence distributed across long trajectories, and TRACES (Li et al., 2026a)

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

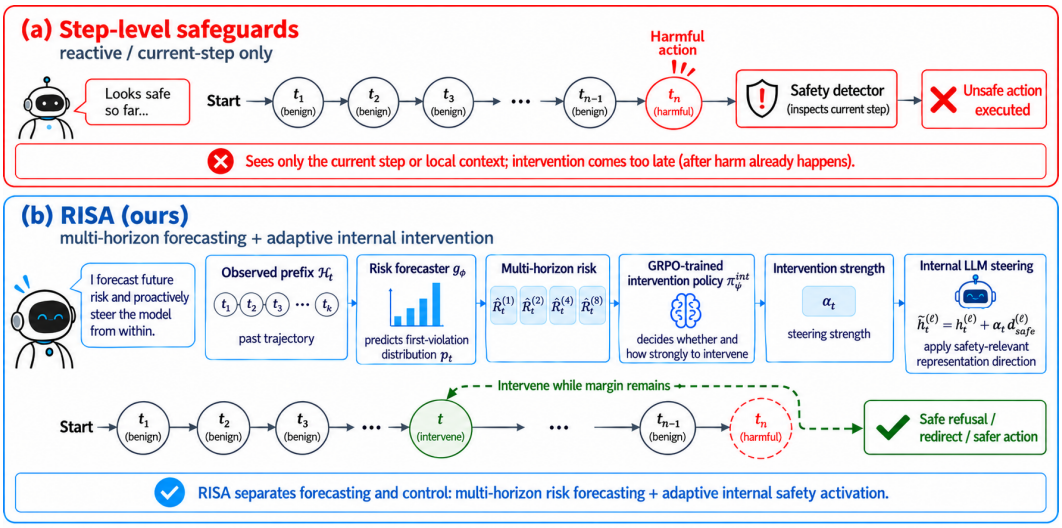

Figure 1: **Overview of RISA for long-horizon agent safety.** Step-level safeguards primarily assess immediate actions, while trajectory-aware approaches incorporate longer interaction histories to detect or anticipate unsafe behavior. RISA combines multi-horizon future-risk prediction and time-to-risk estimation with a learned adaptive intervention policy that controls the strength of an internal safety intervention as trajectory risk evolves.

and DreamGuard (Lin et al., 2026b) explicitly reason about future trajectory risk (See Appendix J for borader discussion). Collectively, these approaches show that partial trajectories can contain predictive signals of unsafe outcomes before a final violating action is executed. However, forecasting that a trajectory is becoming risky and deciding how to intervene are distinct problems.

Two challenges are particularly important for such intervention. First, a single risk estimate does not explicitly distinguish between a violation that is likely to occur in the next step and one that may emerge only after several additional interactions. This distinction matters because the remaining opportunity for corrective action depends on how close the trajectory is to an unsafe event. Second, converting predicted risk into a fixed allow/block decision does not capture how intervention strength should vary as risk evolves. A hard intervention may occur too late when early corrective steering would have been sufficient, while applying strong intervention uniformly can unnecessarily disrupt benign trajectories.

We introduce **RISA** (**R**isk-**I**nformed **S**afety **A**ctivation), a framework that couples future-risk forecasting with adaptive intervention for long-horizon tool-using agents. Figure 1 provides an overview. Given an observed trajectory prefix, RISA estimates the probability of a future unsafe event over multiple prediction horizons and estimates time-to-risk, providing complementary information about the likelihood and temporal urgency of an emerging failure. RISA then uses these signals to control an internal safety intervention rather than mapping risk directly to a fixed allow/block rule. Specifically, a Group Relative Policy Optimization (GRPO) (Shao et al., 2024) trained intervention policy determines when intervention is warranted and how strongly a safety-relevant representation direction is applied to the acting model. Complete trajectories are used during training to construct supervision for future outcomes and intervention timing, whereas inference depends only on quantities predicted from the observed trajectory prefix.

This formulation separates three components of long-horizon safety: *forecasting*, which estimates whether unsafe behavior is likely to emerge over different future horizons; *timing*, which estimates how much intervention margin remains; and *control*, which determines the strength of the resulting safety intervention. We evaluate these components using multi-horizon risk-prediction quality, time-to-risk accuracy, warning lead time, and end-to-end measures of unsafe completion, task utility, and over-refusal. We further compare adaptive intervention with fixed intervention strategies to examine whether using trajectory-dependent risk and urgency can reduce unsafe outcomes while limiting unnecessary intervention on benign trajectories.

Contributions. Our main contributions are:

- We develop a **multi-horizon future-risk formulation** for long-horizon agent safety that distinguishes near-term from longer-horizon unsafe outcomes using the observed trajectory prefix.
- We incorporate **time-to-risk estimation** to complement future-risk probabilities with an explicit signal of temporal urgency and remaining intervention margin.
- We introduce **RISA**, which couples these predictive signals with a GRPO trained intervention policy that adaptively determines when and how strongly to apply a safety-relevant representation intervention.
- We evaluate RISA across multiple long-horizon safety benchmarks and model families, measuring forecasting quality, time-to-risk prediction, warning lead time, unsafe completion, task utility, over-refusal, and the effect of adaptive versus fixed intervention strategies.

2 Problem Formulation

We consider a tool-using language-model agent parameterized by θ and denoted by π_θ. The agent interacts sequentially with an external environment over a finite number of decision steps. At each step $t \in \{1, \dots, T\}$, the agent receives a user or task input $x_t \in \mathcal{X}$, produces an action $a_t \in \mathcal{A}$, and receives an observation $o_t \in \mathcal{O}$ from the environment or an invoked tool, where $\mathcal{X}$, $\mathcal{A}$, and $\mathcal{O}$ denote the input, action, and observation spaces, respectively. The agent policy π_θ defines a conditional distribution over actions given the interaction history. After step t has been executed, we denote the observed trajectory prefix by

$$H_t = (x_1, a_1, o_1, \dots, x_t, a_t, o_t), \tag{1}$$

where H_t contains all information available up to and including step t. When $t = T$, H_T denotes the complete interaction trajectory.

Let $y_t \in 0, 1$ be a binary safety indicator associated with the action executed at step t, where $y_t = 1$ indicates that a_t violates the task-specific safety policy and $y_t = 0$ indicates that no violation is observed at that step. Importantly, $y_t = 0$ does not imply that the trajectory prefix H_t is globally safe. In long-horizon interactions, a sequence of individually permissible actions may progressively move the environment toward a state from which a later unsafe action becomes increasingly likely.

We therefore characterize the safety state of a trajectory prefix using *multi-horizon future risk*. Let

$$\mathcal{K} = \{k_1, \dots, k_M\} \tag{2}$$

denote a predefined set of M positive prediction horizons, where each $k_m \in \mathbb{N}{>}\, 0$ specifies the number of future decision steps over which risk is evaluated. For any horizon $k \in \mathcal{K}$, we define the future-risk function

$$R_t^{(k)} = \Pr(\exists j \in \{1, \dots, k\} : y_{t+j} = 1 \mid H_t), \qquad k \in \mathcal{K}. \tag{3}$$

Here, $R_t^{(k)} \in [0, 1]$ denotes the probability of at least one safety violation within the next k valid steps, conditioned on the observed trajectory prefix H_t. The index j denotes the future-step offset, so y_{t+j} is the safety outcome j steps ahead. If the trajectory terminates before completing k additional steps, only the remaining valid steps are considered.

For the horizon set $\mathcal{K}$, we collect the corresponding risks into the multi-horizon risk vector

$$\boldsymbol{R}_t = \left[R_t^{(k_1)}, \dots, R_t^{(k_M)} \right] \in [0, 1]^M. \tag{4}$$

The vector $\mathbf{R}_t$ describes how the predicted probability of a future violation changes as the prediction horizon increases. For example, a trajectory prefix may have low short-horizon risk $R_t^{(k_1)}$ while exhibiting substantially larger long-horizon risk $R_t^{(k_M)}$, indicating that the current state is not immediately unsafe but is evolving toward a potentially unsafe outcome.

Estimating future risk alone is insufficient for intervention. A safety mechanism must additionally determine *when* an intervention should be activated and *how strongly* it should modify the agent's behavior. We therefore introduce an intervention policy μ_ϕ, parameterized by ϕ, that maps information available from the current trajectory prefix to an intervention strength:

$$\alpha_t = \mu_\phi(H_t, \boldsymbol{R}_t), \qquad \alpha_t \in \mathcal{S}, \tag{5}$$

where α_t denotes the intervention strength at step t and $\mathcal{S}$ denotes the allowable intervention-strength space. We take $\mathcal{S} = [0, 1]$, where $\alpha_t = 0$ corresponds to no intervention and larger values indicate

progressively stronger safety modification. The specific mechanism by which α_t alters the agent is introduced in Section 3.

Objective. Our goal is to learn an intervention mechanism that adapts its activation and strength to the evolving safety state of the trajectory while minimizing unnecessary modification of benign behavior. Complete trajectories are used during training to supervise future-risk estimation, whereas intervention decisions depend only on the observed history H_t. This requires jointly estimating multi-horizon future risk and using these estimates to determine when and how strongly to intervene. We next introduce RISA, which implements this through multi-horizon risk prediction, safety-relevant representation steering, and a learned intervention policy.

3 RISA: RISK-INFORMED SAFETY ACTIVATION

We now present RISA, a risk-informed intervention framework that combines multi-horizon trajectory risk prediction with adaptive representation steering. We first describe how future safety risk is estimated from the observed trajectory prefix, then construct a safety-relevant representation direction, and finally introduce the adaptive intervention policy and its trajectory-aware training objective for learning when and how strongly to intervene.

3.1 MULTI-HORIZON RISK PREDICTION

Given the observed trajectory prefix H_t, RISA estimates both whether a future safety violation is likely to occur and how far ahead it may emerge. Rather than learning an independent predictor for each horizon, we use a shared predictor to model the distribution over the step at which the first future violation occurs. This follows the multi-horizon forecasting principle of jointly modeling correlated future outcomes common in finance (Lim et al., 2021; Zhang & Zohren, 2021).
Let $k_M = \max \mathcal{K}$ denote the maximum prediction horizon. For a prefix H_t, let

$$d_t = \min \{j \geq 1 : y_{t+j} = 1\} \tag{6}$$

denote the offset to the first future safety violation, whenever such a violation occurs.
Given H_t, the predictor g_ϕ outputs a categorical distribution

$$\boldsymbol{p}_t = g_\phi(H_t) = \left[p_t^{(1)}, \ldots, p_t^{(k_M)}, p_t^{(\varnothing)}\right], \tag{7}$$

where

$$p_t^{(j)} = \Pr(d_t = j \mid H_t) \tag{8}$$

denotes the probability that the first future safety violation occurs exactly j interaction steps after the current prefix. The additional outcome $p_t^{(\varnothing)}$ denotes the probability that no violation occurs within the next k_M valid steps. The distribution satisfies

$$\sum_{j=1}^{k_M} p_t^{(j)} + p_t^{(\varnothing)} = 1. \tag{9}$$

The predicted risk for any horizon $k \in \mathcal{K}$ is obtained by accumulating the probability of a first violation occurring within that horizon:

$$\widehat{R}_t^{(k)} = \sum_{j=1}^{k} p_t^{(j)}, \qquad k \in \mathcal{K}. \tag{10}$$

The resulting multi-horizon risk vector is

$$\widehat{\boldsymbol{R}}_t = \left[\widehat{R}_t^{(k_1)}, \ldots, \widehat{R}_t^{(k_M)}\right]. \tag{11}$$

Because these risks are cumulative probabilities derived from the same first-violation distribution, they are naturally non-decreasing with the prediction horizon:

$$\widehat{R}_t^{(k_1)} \leq \widehat{R}_t^{(k_2)} \leq \cdots \leq \widehat{R}_t^{(k_M)}, \qquad k_1 < k_2 < \cdots < k_M. \tag{12}$$

Thus, no additional monotonicity regularization is required.

During training, complete trajectories provide supervision for the first-violation offset. For each prefix H_t, we construct the categorical target

$$\widetilde{d}_t = \begin{cases} d_t, & \text{if } d_t \le k_M, \\ \varnothing, & \text{otherwise}, \end{cases} \tag{13}$$

where $\varnothing$ also applies when no future violation occurs within the next k_M valid steps, including cases in which the trajectory terminates before a violation is observed.

The predictor is trained by minimizing the negative log-likelihood of the observed target:

$$\mathcal{L}_{\text{risk}} = -\frac{1}{|\mathcal{P}|} \sum_{t \in \mathcal{P}} \log p_t^{(\widetilde{d}_t)}, \tag{14}$$

where $\mathcal{P}$ denotes the set of training prefixes. Future trajectory information is used only to construct the supervision target $\widetilde{d}_t$; the predictor itself receives only H_t. At inference time, $\widehat{\boldsymbol{R}}_t$ is therefore estimated solely from the observed trajectory prefix, without access to future actions, safety outcomes, or the eventual trajectory length.

3.2 Safety-Relevant Representation Direction

Multi-horizon risk prediction determines when intervention may be required, but the intervention itself should help the acting model make a safer decision without replacing its underlying capabilities. RISA therefore derives a safety-relevant direction directly from the model's hidden representations. The purpose of this direction is to amplify latent features associated with safer trajectory states, thereby encouraging the model to make greater use of safety-relevant information already encoded in its internal representation space.

For an observed trajectory prefix H_t, let

$$h_t^{(\ell)} = f_\theta^{(\ell)}(H_t) \in \mathbb{R}^d, \tag{15}$$

where $f_\theta^{(\ell)}$ denotes the mapping induced by the acting model π_θ from the input trajectory prefix to its hidden representation at transformer layer ℓ. Specifically, $h_t^{(\ell)}$ is the hidden state of the final token of H_t, immediately before generation of the next action.

Using the first-violation targets defined in Section 3.1, we partition training prefixes into future-risk and safe groups:

$$\mathcal{P}_{\text{risk}} = \left\{ t : \widetilde{d}_t \in \{1, \dots, k_M\} \right\}, \tag{16}$$

$$\mathcal{P}_{\text{safe}} = \left\{ t : \widetilde{d}_t = \varnothing \right\}. \tag{17}$$

Here, $\mathcal{P}_{\text{risk}}$ contains prefixes followed by a safety violation within the maximum forecasting horizon k_M, whereas $\mathcal{P}_{\text{safe}}$ contains prefixes for which no violation occurs within the forecasting window.

For each layer ℓ, we compute the mean hidden representation of the two groups:

$$\mu_{\text{risk}}^{(\ell)} = \frac{1}{|\mathcal{P}_{\text{risk}}|} \sum_{t \in \mathcal{P}_{\text{risk}}} h_t^{(\ell)}, \qquad \mu_{\text{safe}}^{(\ell)} = \frac{1}{|\mathcal{P}_{\text{safe}}|} \sum_{t \in \mathcal{P}_{\text{safe}}} h_t^{(\ell)}. \tag{18}$$

We then define the safety-relevant direction as the normalized difference between the safe and future-risk representation centroids:

$$d_{\text{safe}}^{(\ell)} = \frac{\mu_{\text{safe}}^{(\ell)} - \mu_{\text{risk}}^{(\ell)}}{\left\| \mu_{\text{safe}}^{(\ell)} - \mu_{\text{risk}}^{(\ell)} \right\|_2}. \tag{19}$$

The direction $d_{\text{safe}}^{(\ell)}$ captures a linear contrast between hidden states associated with trajectories that remain safe and those that later lead to a violation. Its orientation is chosen toward the safe representation centroid, so positive steering along this direction is intended to strengthen internal features associated with safer decision-making. Rather than injecting an external safety rule into the model, RISA uses this direction to modulate safety-relevant structure already present in the model's hidden space. Normalization decouples the direction from the intervention magnitude, allowing the adaptive intervention policy to control steering strength independently through α_t.

The safety direction is estimated exclusively from the training split and remains fixed during intervention-policy learning and inference. Future trajectory outcomes are used only to construct

the offline safe and future-risk groups; at inference time, applying $d_{\text{safe}}^{(\ell)}$ requires only the current hidden representation and does not rely on future trajectory information.

3.3 Risk-Informed Adaptive Intervention

The multi-horizon risk profile indicates how urgently a trajectory may require intervention, while the safety direction provides a mechanism for activating safety-relevant structure in the acting model. RISA connects these components through a learned intervention policy that determines both *whether* to intervene and *how strongly* to steer the model.

Given the predicted risk vector $\widehat{\boldsymbol{R}}_t$, the intervention policy π_ψ^{int} selects a steering strength

$$\alpha_t \sim \pi_\psi^{\text{int}}\left(\cdot \mid \widehat{\boldsymbol{R}}_t\right), \qquad \alpha_t \in [0,1], \tag{20}$$

where α_t denotes the allowable intervention strengths and includes $\alpha_t = 0$ as the no-intervention option. Thus, the same policy determines whether intervention is necessary and, when activated, its magnitude. For a selected strength α_t, RISA modifies the hidden representation at the intervention layer ℓ as

$$\widetilde{h}_t^{(\ell)} = h_t^{(\ell)} + \alpha_t d_{\text{safe}}^{(\ell)}. \tag{21}$$

The modified representation is then used by the acting model to generate the next action. Since $d_{\text{safe}}^{(\ell)}$ is normalized, α_t directly controls the magnitude of the safety-relevant modification: $\alpha_t = 0$ leaves the representation unchanged, while larger values increasingly emphasize the learned safety direction.

Trajectory-aware intervention reward. The intervention policy should favor steering that reduces future safety risk while preserving task performance and avoiding unnecessary modification of benign trajectories. During training, at a given trajectory prefix H_t, we consider multiple possible intervention actions, where each intervention action corresponds to a different steering strength.

Let $\alpha_t^{(i)}$ denote the steering strength of the i-th candidate intervention applied at the same current prefix H_t. Each candidate therefore represents an alternative intervention choice, rather than a different trajectory state. Applying $\alpha_t^{(i)}$ changes the model representation, after which the model generates and executes the next action, producing the resulting continuation $H_{t+1}^{(i)}$. The corresponding post-intervention risk estimate at horizon k is denoted by $\widehat{R}_{t+1}^{(k,i)}$. The corresponding post-intervention risk estimate is $\widehat{R}_{t+1}^{(k,i)}$. To isolate the effect of intervention, we compare each candidate with the no-intervention continuation $\alpha_t^{(0)} = 0$ from the same prefix:

$$\Delta_t^{(i)} = \sum_{k \in \mathcal{K}} w_k \left(\widehat{R}_{t+1}^{(k,0)} - \widehat{R}_{t+1}^{(k,i)}\right), \tag{22}$$

where $w_k \geq 0$ controls the contribution of horizon k and $\sum_{k\in\mathcal{K}} w_k = 1$. Positive $\Delta_t^{(i)}$ therefore indicates that intervention reduces predicted future risk relative to leaving the model unmodified.

Complete training trajectories additionally provide supervision about how far the current prefix lies from the first future violation. Using the first-violation target $\widetilde{d}_t$ from Section 3.1, we define the normalized lead-time factor

$$\eta_t = \begin{cases} \dfrac{\widetilde{d}_t}{k_M}, & \widetilde{d}_t \in \{1, \ldots, k_M\}, \\ 0, & \widetilde{d}_t = \varnothing. \end{cases} \tag{23}$$

Larger η_t corresponds to a violation that lies farther in the future, allowing successful interventions made earlier in the trajectory to receive additional credit.

The reward assigned to candidate intervention i is

$$q_t^{(i)} = \lambda_r \left(1 + \lambda_e \eta_t\right) \Delta_t^{(i)} + \lambda_s r_{\text{safe}}^{(i)} + \lambda_u r_{\text{task}}^{(i)} - \lambda_c \left(\alpha_t^{(i)}\right)^2, \tag{24}$$

where $r_{\text{safe}}^{(i)}$ measures the safety outcome of the resulting continuation, $r_{\text{task}}^{(i)}$ measures task utility, and $\lambda_r, \lambda_e, \lambda_s, \lambda_u, \lambda_c \geq 0$ control the relative contributions of risk reduction, early intervention, safety, utility, and intervention cost. The intervention penalty discourages unnecessarily strong steering, while the task-utility term discourages safety gains obtained through indiscriminate refusal or degradation of useful behavior.

Group-relative policy learning. We optimize π_ψ^{int} using GRPO (Shao et al., 2024). For each trajectory prefix, we evaluate a group of G candidate steering strengths

$$\alpha_t^{(1)}, \ldots, \alpha_t^{(G)} \sim \pi_{\psi_{\text{old}}}^{\text{int}} \left(\cdot \mid \widehat{\boldsymbol{R}}_t \right), \tag{25}$$

where ψ_{old} denotes the intervention-policy parameters before the current GRPO update and is treated as fixed while optimizing the updated parameters ψ. The no-intervention action $\alpha_t^{(0)} = 0$ is included as a reference candidate. Each candidate is evaluated using Eq. (24). We compute its group relative advantage as

$$A_t^{(i)} = \frac{q_t^{(i)} - \overline{q}_t}{\sigma_t + \epsilon}, \tag{26}$$

where $\overline{q}_t$ and σ_t are the mean and standard deviation of rewards within the candidate group and $\epsilon > 0$ is a numerical-stability constant.

The intervention policy is updated using the clipped objective

$$\mathcal{L}_{\text{int}} = -\mathbb{E}_{t,i} \left[\min \left(\rho_t^{(i)}(\psi) A_t^{(i)}, \text{clip} \left(\rho_t^{(i)}(\psi), 1 - \epsilon_{\text{clip}}, 1 + \epsilon_{\text{clip}} \right) A_t^{(i)} \right) \right], \tag{27}$$

where

$$\rho_t^{(i)}(\psi) = \frac{\pi_\psi^{\text{int}} \left(\alpha_t^{(i)} \mid \widehat{\boldsymbol{R}}_t \right)}{\pi_{\psi_{\text{old}}}^{\text{int}} \left(\alpha_t^{(i)} \mid \widehat{\boldsymbol{R}}_t \right)}. \tag{28}$$

This group-relative formulation directly compares alternative intervention strengths under the same risk state, encouraging the policy to select the smallest intervention that sufficiently reduces future risk while preserving task utility. During intervention-policy learning, the acting model π_θ, multi-horizon risk predictor g_ϕ, and safety direction $d_{\text{safe}}^{(\ell)}$ remain fixed; only the intervention-policy parameters ψ are updated. The lead-time target η_t and rollout outcomes are used only as training supervision and are not available to the policy at inference time. Complete pseudocode for risk-predictor training, safety-direction construction, and intervention-policy optimization is provided in Appendix A.

4 Experiments

We evaluate RISA around three questions: **(1)** Can multi-horizon risk forecasting identify emerging safety failures before the violating action becomes immediate? **(2)** Does risk-conditioned intervention improve end-to-end safety without achieving this improvement through excessive refusal? **(3)** Are multi-horizon conditioning and adaptive intervention necessary for this trade-off? We additionally measure inference overhead and evaluate generalization across model families.

4.1 Experimental Setup

Benchmarks and models. We evaluate on three long-horizon agent-safety benchmarks: **BLINDSPOT** (Asif et al., 2026), **MT-AgentRisk** (Li et al., 2026b), and **LongDS-Bench** (Xu et al., 2026). All contain multi-step interactions in which locally plausible actions can accumulate toward a later safety violation. Results for **Qwen3-4B** and **Gemma-2-9B** are reported in Section 4.3. Results for **LLaMA-8B**, **Qwen2.5-14B** and additional model scales are deferred to Appendix C. For each backbone, the risk predictor, safety direction, and intervention policy are trained independently; no learned intervention components are transferred across model families.

Baselines. We compare against the unprotected **Base Agent**; the step-level safeguard **ToolSafe** (Mou et al., 2026); trajectory-aware **TRACE** (Hong et al., 2026); proactive trajectory-risk method **TRACES** (Li et al., 2026a); and rubric based agent safeguard **Safety Rubric** (Loye et al., 2026). For controlled ablations, we additionally compare against fixed steering, threshold-triggered steering, and a single-horizon variant of RISA.

Evaluation. For risk forecasting, we report area under the receiver operating characteristic curve (**AUROC**) and area under the precision-recall curve (**AUPRC**) at each horizon. For end-to-end behavior, we report Unsafe Completion Rate (**UCR**), Safe Completion Rate (**SCR**), Over-Refusal Rate (**ORR**), and task utility. We additionally report warning lead time in intervention ablations and per-step inference overhead. Formal metric definitions, training details, hyperparameters, and the construction of first-violation targets are provided in Appendix B.

Table 2: Safety and utility across long-horizon agent-safety benchmarks for Qwen3-4B and Gemma-2-9B. Base Agent provides the unsafeguarded utility reference. Best safeguarded result for each model and metric is shown in bold.

		Qwen3-4B				Gemma-2-9B			
Benchmark	**Method**	**UCR↓**	**Utility↑**	**ORR↓**	**SCR↑**	**UCR↓**	**Utility↑**	**ORR↓**	**SCR↑**
BLINDSPOT	Base Agent	0.432	0.842	0.057	0.514	0.447	0.831	0.061	0.501
	ToolSafe	0.580	0.605	0.135	0.610	0.600	0.585	0.150	0.595
	TRACE	0.430	0.700	0.240	0.655	0.460	0.675	0.255	0.640
	TRACES	0.385	0.720	0.175	0.690	0.405	0.705	0.185	0.680
	SafetyRubric	0.375	0.735	0.170	0.700	0.400	0.715	0.180	0.690
	RISA	**0.300**	**0.775**	**0.120**	**0.775**	**0.320**	**0.765**	**0.130**	**0.760**
MT-AgentRisk	Base Agent	0.394	0.858	0.051	0.553	0.408	0.846	0.055	0.541
	ToolSafe	0.272	0.792	0.126	0.649	0.286	0.782	0.133	0.637
	TRACE	0.236	0.808	0.106	0.678	0.248	0.798	0.113	0.667
	TRACES	0.205	0.799	0.114	0.704	0.216	0.790	0.120	0.642
	SafetyRubric	0.240	0.781	0.132	0.693	0.187	0.770	0.139	0.710
	RISA	**0.128**	**0.831**	**0.079**	**0.723**	**0.139**	**0.821**	**0.085**	**0.760**
LongDS-Bench	Base Agent	0.361	0.847	0.049	0.588	0.374	0.838	0.052	0.575
	ToolSafe	0.245	0.787	0.116	0.671	0.258	0.777	0.122	0.660
	TRACE	0.283	0.801	0.101	0.646	0.227	0.783	0.107	0.685
	TRACES	0.189	0.795	0.108	0.717	0.199	0.736	0.114	0.706
	SafetyRubric	0.161	0.773	0.129	0.701	0.171	0.764	0.120	0.650
	RISA	**0.118**	**0.824**	**0.075**	**0.782**	**0.127**	**0.795**	**0.081**	**0.770**

4.2 RISA Anticipates Violations Before They Become Immediate

We first evaluate the shared risk predictor on held-out trajectory prefixes. Unlike training independent classifiers for each horizon, RISA predicts a single distribution over first-violation time and derives each $\widehat{R}_t^{(k)}$ by accumulating the corresponding probability mass according to Eq. (10).

Forecasting becomes progressively harder as the prediction horizon increases, but predictive performance degrades smoothly rather than collapsing. AUROC decreases from 0.914 at one step to 0.801 at eight steps, while AUPRC decreases from 0.846 to 0.689. Thus, trajectory prefixes contain detectable safety-relevant signals several interaction steps before a violation becomes immediate.

Table 1: Multi-horizon future-risk forecasting on held-out trajectory prefixes for Qwen3-4B on the BLINDSPOT benchmark.

Horizon	**AUROC↑**	**AUPRC↑**
$k = 1$	0.914	0.846
$k = 2$	0.889	0.812
$k = 4$	0.853	0.758
$k = 8$	0.801	0.689

Because the predictor models first-violation time directly, we also evaluate its temporal forecasting accuracy using negative log-likelihood (**NLL**) and mean absolute error (**MAE**). These results, together with horizon-wise calibration analyses, are reported in Appendix D.

4.3 End-to-End Safety and Utility

We next evaluate whether early risk forecasting translates into safer agent behavior. This evaluation considers safety and utility jointly: unsafe completion can be reduced trivially by refusing more often, so UCR alone is insufficient to characterize a useful safeguard.

RISA achieves the lowest unsafe completion rate across all three benchmarks for both Qwen3-4B and Gemma-2-9B, while retaining the highest task utility and lowest over-refusal among the evaluated safeguards. On BLINDSPOT with Qwen3-4B, RISA reduces unsafe completion from 0.432 for the Base Agent to 0.300, while retaining a utility of 0.775. On Gemma-2-9B, unsafe completion decreases from 0.447 to 0.320, with utility remaining at 0.765.

The same trend holds on MT-AgentRisk and LongDS-Bench. For Qwen3-4B, RISA achieves UCRs of 0.128 and 0.118, respectively, while maintaining utilities of 0.831 and 0.824. For Gemma-2-9B, RISA achieves UCRs of 0.139 and 0.127, with corresponding utilities of 0.821 and 0.795. Across all benchmark–model combinations, RISA also yields the lowest over-refusal rate and highest safe completion rate among the evaluated safeguards.

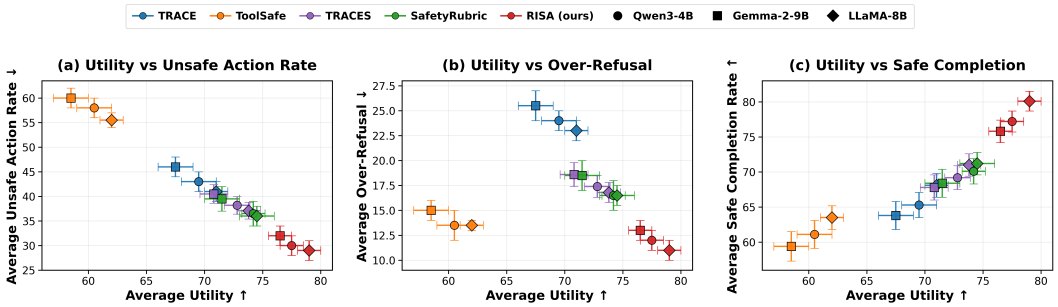

Figure 2: **Safety-utility trade-off across model families.** Colors denote safety methods and marker shapes denote acting-model backbones; error bars show variation across repeated runs. RISA consistently shifts the operating point toward lower unsafe completion while retaining higher task utility compared with the evaluated safeguards.

Although the unsafeguarded Base Agent provides the highest raw utility, RISA retains substantially more utility than the other safety methods while providing stronger safety improvements. Results for LLaMA-8B and additional model scales are reported in Appendix C.

Figure 2 further illustrates this safety–utility trade-off across model families. RISA consistently occupies a region with lower unsafe completion and over-refusal, and higher safe completion, while retaining comparatively high task utility. The same qualitative pattern is observed across Qwen3-4B, Gemma-2-9B, and the additional model families evaluated in Appendix C. Full benchmark-level results for each backbone are provided in the appendix.

4.4 Why Does Adaptive Multi-Horizon Intervention Help?

We isolate the two central design choices in RISA: adapting intervention strength to the current risk state and using the full multi-horizon risk profile rather than a single scalar threshold. Applying the same steering magnitude at every step reduces unsafe behavior but substantially harms utility and increases over-refusal.

Table 3: Ablation of the intervention strategy for LLaMA-8B on the BLINDSPOT. (Lead = Warning Lead Time)

Strategy	UCR↓	Utility↑	ORR↓	Lead↑
Fixed Steering	0.251	0.742	0.183	2.21
Threshold Steering	0.163	0.759	0.154	2.48
Single-Horizon RISA	0.141	0.793	0.108	2.73
RISA w/o Lead Reward	0.133	0.805	0.091	1.94
RISA	**0.106**	**0.820**	**0.074**	**3.29**

Conditioning intervention on a risk threshold improves this trade-off, but remains substantially worse than learned adaptive intervention. Using only the longest-horizon risk also degrades performance: Single-Horizon RISA increases UCR from 0.106 to 0.141 and ORR from 0.074 to 0.108 while reducing utility from 0.820 to 0.793. This provides direct evidence that the temporal shape of the risk profile contains useful control information beyond a single long-horizon probability.

The lead-time reward affects a different aspect of behavior. Removing it reduces warning lead time from 3.29 to 1.94 steps, with a smaller change in the remaining metrics. This supports its intended role: encouraging earlier intervention rather than merely increasing intervention strength.

Representation direction. We additionally verify that the intervention is not explained by arbitrary hidden-state perturbation. Under matched perturbation magnitudes, steering along $+d_{\text{safe}}$ reduces UCR from 0.432 to 0.276, compared with 0.407 for a random direction, whereas reversing the direction increases UCR to 0.468. Layer-wise separability, matched-direction controls, and intervention strength as a function of the multi-horizon risk profile are reported in Appendix E.

5 Conclusion

We introduced RISA, a risk-informed safety framework for long-horizon tool-using agents that combines multi-horizon future-risk forecasting with adaptive representation steering. By explicitly separating risk prediction from intervention control, RISA can anticipate emerging unsafe trajectories and adjust intervention strength according to their evolving risk profile. Across multiple benchmarks and model families, RISA reduces unsafe completion while preserving task utility and limiting over-refusal.

AI USE STATEMENT

Claude Opus 4.8 (Anthropic) and GPT-5.6 Sol (OpenAI) were used for editing, proofreading, resolving LaTeX formatting issues, and to assist with code development for the implementation of the proposed methods. All AI-assisted text and code were reviewed, verified, and, where necessary, modified by the authors.

REFERENCES

Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, J. Zico Kolter, Matt Fredrikson, Yarin Gal, and Xander Davies. Agentharm: A benchmark for measuring harmfulness of llm agents. In *International Conference on Learning Representations*, 2025.

Sadia Asif and Mohammad Mohammadi Amiri. Information-theoretic privacy control for sequential multi-agent llm systems. *arXiv preprint arXiv:2603.05520*, 2026. doi: 10.48550/arXiv.2603.05520. URL https://doi.org/10.48550/arXiv.2603.05520.

Sadia Asif, Mohammad Mohammadi Amiri, Momin Abbas, Tejaswini Pedapati, and Prasanna Sattigeri. Blindspot: A benchmark for safety and refusal calibration in long-horizon tool-using agents, 2026. URL https://arxiv.org/abs/2609.16305.

Roberto Campbell, Momin Abbass, Muneeza Azmat, Michal Ulewicz, Raya Horesh, Kristjan Greenewald, Rogério Abreu de Paula, and Nathalie Baracaldo. An efficient and modular framework for targeted harm mitigation in llms, 2026. URL https://arxiv.org/abs/2609.13624.

Zhaorun Chen, Mintong Kang, and Bo Li. Shieldagent: Shielding agents via verifiable safety policy reasoning. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=DkRYImuQA9.

Sahana Chennabasappa, Cyrus Nikolaidis, Daniel Song, David Molnar, Stephanie Ding, Shengye Wan, Spencer Whitman, Lauren Deason, Nicholas Doucette, Abraham Montilla, et al. Llamafirewall: An open source guardrail system for building secure ai agents. *arXiv preprint arXiv:2505.03574*, 2025.

Edoardo Debenedetti, Jie Zhang, Mislav Balunovic, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. Agentdojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Guanting Dong, Xiaoshuai Song, Yuyang Hu, Jiajie Jin, Chenghao Zhang, Yifei Chen, Xiaoxi Li, Huaying Yuan, Xinyu Yang, Tongyu Wen, Jiejun Tan, Hongjin Qian, Shijue Huang, Junting Lu, Zhenyu Li, Wanjun Zhong, Yutao Zhu, Tat-Seng Chua, Zhicheng Dou, and Ji-Rong Wen. Towards long-horizon agents: A survey. *Preprints*, July 2026. doi: 10.20944/preprints202607.1328.v1. URL https://doi.org/10.20944/preprints202607.1328.v1.

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.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*, 2023.

Muhammad Zeeshan Karamat, Sadia Asif, Haris Khan, Muhammad Shaheer, Shumaila Asif, Rajesh Upadhayaya, Christiana Chamon, and Mohammad Mohammadi Amiri. Privacy, security, and data leakage concerns in multi-agent ai systems: A literature review. 2026. doi: 10.5281/zenodo.19908321. URL https://doi.org/10.5281/zenodo.19908321.

Jiaqian Li, Yanshu Li, Ruixiang Tang, Boxuan Zhang, and Kuan-Hao Huang. TRACES: Proactive safety auditing for multi-turn llm agents via trajectory-state modeling. *arXiv preprint arXiv:2605.27690*, 2026a. URL https://arxiv.org/abs/2605.27690.

Xu Li, Simon Yu, Minzhou Pan, Yiyou Sun, Bo Li, Dawn Song, Xue Lin, and Weiyan Shi. Unsafer in many turns: Benchmarking and defending multi-turn safety risks in tool-using agents. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026b.

Bryan Lim, Sercan O. Arik, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4): 1748–1764, 2021. doi: 10.1016/j.ijforecast.2021.03.012.

Shi Lin, Peng Qian, Dinghao Liu, Renjie Sun, Sifan Wu, Dezhang Kong, Chenpei Wang, and Xun Wang. Forecasting trajectory-level safety risks in black-box multi-turn interactions. *arXiv preprint arXiv:2607.26820*, 2026a.

Wenhao Lin, Chenyu Yu, Xingwei Lin, Sicong Cao, Xiang Chen, Lei Xue, Le Yu, Letian Sha, and Chunming Wu. DreamGuard: Efficient runtime guardrail for llm agents via risk-aware world model. *arXiv preprint arXiv:2608.05695*, 2026b. URL https://arxiv.org/abs/2608.05695.

Dongrui Liu, Qihan Ren, Chen Qian, Shuai Shao, Yuejin Xie, Yu Li, Zhonghao Yang, Haoyu Luo, Peng Wang, Qingyu Liu, Binxin Hu, Ling Tang, Jilin Mei, Dadi Guo, Leitao Yuan, Junyao Yang, Guanxu Chen, Qihao Lin, Yi Yu, Bo Zhang, Jiaxuan Guo, Jie Zhang, Wenqi Shao, Huiqi Deng, Zhiheng Xi, Wenjie Wang, Wenxuan Wang, Wen Shen, Zhikai Chen, Haoyu Xie, Jialing Tao, Juntao Dai, Jiaming Ji, Zhongjie Ba, Linfeng Zhang, Yong Liu, Quanshi Zhang, Lei Zhu, Zhihua Wei, Hui Xue, Chaochao Lu, Jing Shao, and Xia Hu. Agentdog: A diagnostic guardrail framework for ai agent safety and security, 2026. URL https://arxiv.org/abs/2601.18491.

Xian Qi Loye, Qinglin Su, Zhexin Zhang, Shiyao Cui, Qi Zhu, Fei Mi, Hongning Wang, and Minlie Huang. Rubas: Rubric-based reinforcement learning for agent safety. *arXiv preprint arXiv:2606.04051*, 2026.

Yutao Mou, Zhangchi Xue, Lijun Li, Peiyang Liu, Shikun Zhang, Wei Ye, and Jing Shao. ToolSafe: Enhancing tool invocation safety of LLM-based agents via proactive step-level guardrail and feedback. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 37125–37153, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1850. URL https://aclanthology.org/2026.findings-acl.1850/.

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

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*.

Zhen Xiang, Linzhi Zheng, Yanjie Li, Junyuan Hong, Qinbin Li, Han Xie, Jiawei Zhang, Zidi Xiong, Chulin Xie, Carl Yang, Dawn Song, and Bo Li. Guardagent: Safeguard LLM agents via knowledge-enabled reasoning. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=2nBcjCZrrP.

Yuan Xiong, Linji Hao, Shizhu He, Yequan Wang, and Lijun Li. Janus: Foreseeing latent risk for long-horizon agent safety. *arXiv preprint arXiv:2607.19913*, 2026.

Kewei Xu, Xiaoben Lu, Shuofei Qiao, Zihan Ding, Haoming Xu, Lei Liang, and Ningyu Zhang. Longds-bench: On the failure of long-horizon agentic data analysis, 2026. URL https://arxiv.org/abs/2605.30434.

Wenjun Zeng, Yuchi Liu, Ryan Mullins, Ludovic Peran, Joe Fernandez, Hamza Harkous, Karthik Narasimhan, Drew Proud, Piyush Kumar, Bhaktipriya Radharapu, Olivia Sturman, and Oscar Wahltinez. Shieldgemma: Generative ai content moderation based on gemma, 2024. URL https://arxiv.org/abs/2407.21772.

Boxuan Zhang, Jianing Zhu, Zeru Shi, Dongfang Liu, and Ruixiang Tang. Agentforesight: Online auditing for early failure prediction in multi-agent systems. *arXiv preprint arXiv:2605.08715*, 2026.

Zihao Zhang and Stefan Zohren. Multi-horizon forecasting for limit order books: Novel deep learning approaches and hardware acceleration using intelligent processing units. *arXiv preprint arXiv:2105.10430*, 2021.

A Algorithmic Details

This section summarizes the three-stage training procedure used by RISA. First, complete training trajectories are converted into trajectory prefixes with categorical first-violation supervision and used to train the shared multi-horizon risk predictor, as summarized in Algorithm 1. Second, training prefixes are partitioned into future-risk and safe groups to construct a fixed safety-relevant representation direction, described in Algorithm 2. Finally, with the acting model, risk predictor, and safety direction frozen, the intervention policy is optimized using group-relative policy learning, as detailed in Algorithm 3. We index an intervention decision by the current prefix H_t; therefore, α_t is selected after observing H_t and is applied before generating the next action.

Algorithm 1 Training the Multi-Horizon Risk Predictor

Require: Training trajectories $\mathcal{D}$ with step-level safety labels $\{y_t\}$; horizon set $\mathcal{K}$; risk predictor g_ϕ; number of optimization epochs E_{risk}
Ensure: Trained risk predictor g_ϕ

$k_M \leftarrow \max \mathcal{K}$
Initialize prefix set $\mathcal{P} \leftarrow \emptyset$
for each trajectory $H_T \in \mathcal{D}$ of length T **do**
  for $t = 1, \ldots, T-1$ **do**
    Construct $H_t = (x_1, a_1, o_1, \ldots, x_t, a_t, o_t)$
    $\mathcal{J}_t \leftarrow \{j : 1 \le j \le \min(k_M, T-t),\ y_{t+j} = 1\}$
    if $\mathcal{J}_t \neq \emptyset$ **then**
      $\widetilde{d}_t \leftarrow \min \mathcal{J}_t$
    else
      $\widetilde{d}_t \leftarrow \varnothing$
    end if
    Add $(H_t, \widetilde{d}_t)$ to $\mathcal{P}$
  end for
end for
for $e = 1, \ldots, E_{\text{risk}}$ **do**
  Sample minibatch $\mathcal{B} \subset \mathcal{P}$
  for each $(H_t, \widetilde{d}_t) \in \mathcal{B}$ **do**
    $\boldsymbol{p}_t \leftarrow g_\phi(H_t)$ ▷ $k_M + 1$ categorical outcomes
  end for
  $\mathcal{L}_{\text{risk}} \leftarrow -\dfrac{1}{|\mathcal{B}|} \displaystyle\sum_{(H_t, \widetilde{d}_t) \in \mathcal{B}} \log p_t^{(\widetilde{d}_t)}$
  Update ϕ by minimizing $\mathcal{L}_{\text{risk}}$
end for
return g_ϕ

The safety direction is constructed only from the training split. Prefixes with a first violation inside the maximum prediction horizon are treated as future-risk prefixes, while $\widetilde{d}_t = \varnothing$ denotes prefixes for which no violation is observed within the valid forecasting window. The direction is computed independently for each candidate intervention layer, and the selected direction remains fixed during intervention-policy training and inference. Algorithm 2 summarizes the construction of the safety-relevant representation direction. The intervention policy is trained after freezing π_θ, g_ϕ, and $d_{\text{safe}}^{(\ell)}$. For each prefix, the policy samples G intervention strengths from $[0, 1]$. A separate no-intervention continuation, $\alpha_t^{(0)} = 0$, is used only as a matched reference for measuring predicted risk reduction and is not included in the G samples used for group-relative normalization. Complete-trajectory information is used only to construct the offline training reward; during inference, the policy receives

Algorithm 2 Constructing the Safety-Relevant Representation Direction

Require: Training prefixes $\mathcal{P}$ with targets $\widetilde{d}_t$; acting model π_θ; intervention layer ℓ; maximum horizon k_M

Ensure: Safety direction $d^{(\ell)}_{\text{safe}}$

$\mathcal{P}_{\text{risk}} \leftarrow \emptyset$
$\mathcal{P}_{\text{safe}} \leftarrow \emptyset$
for each $(H_t, \widetilde{d}_t) \in \mathcal{P}$ **do**
  if $\widetilde{d}_t \in \{1, \dots, k_M\}$ **then**
    Add t to $\mathcal{P}_{\text{risk}}$
  else
    Add t to $\mathcal{P}_{\text{safe}}$
  end if
end for
for each $t \in \mathcal{P}_{\text{risk}} \cup \mathcal{P}_{\text{safe}}$ **do**
  $h^{(\ell)}_t \leftarrow f^{(\ell)}_\theta(H_t)$ ▷ final-token hidden state
end for
$\mu^{(\ell)}_{\text{risk}} \leftarrow \dfrac{1}{|\mathcal{P}_{\text{risk}}|} \displaystyle\sum_{t \in \mathcal{P}_{\text{risk}}} h^{(\ell)}_t$
$\mu^{(\ell)}_{\text{safe}} \leftarrow \dfrac{1}{|\mathcal{P}_{\text{safe}}|} \displaystyle\sum_{t \in \mathcal{P}_{\text{safe}}} h^{(\ell)}_t$
$v^{(\ell)} \leftarrow \mu^{(\ell)}_{\text{safe}} - \mu^{(\ell)}_{\text{risk}}$
$d^{(\ell)}_{\text{safe}} \leftarrow \dfrac{v^{(\ell)}}{\|v^{(\ell)}\|_2}$
return $d^{(\ell)}_{\text{safe}}$

only $\widehat{\boldsymbol{R}}_t$ computed from the observed prefix. Algorithm 3 summarizes the complete intervention-policy training procedure.

B Experimental Details

This section now provides additional implementation and evaluation details for the experiments in Section 4. We describe trajectory-prefix construction, the three-stage RISA training procedure, backbone-specific training, baseline and ablation configurations, and evaluation metrics.

B.1 Trajectory Prefix Construction

Each completed trajectory H_T is decomposed into observed prefixes H_t as defined in Section 2. For risk-prediction and intervention training, we use prefixes with $t \in \{1, \dots, T-1\}$ so that each prefix has at least one possible future decision step. For every prefix, we construct the categorical first-violation target $\widetilde{d}_t$ using Eq. (13). Unless otherwise stated, we use

$$\mathcal{K} = \{1, 2, 4, 8\}, \qquad k_M = \max \mathcal{K} = 8.$$

For each horizon $k \in \mathcal{K}$, we evaluate $\widehat{R}^{(k)}_t$ against the binary event $\mathbb{I}\left[\widetilde{d}_t \in \{1, \dots, k\}\right]$. Thus, all horizon-specific evaluation targets are derived from the same categorical first-violation supervision used to train the shared predictor; no independent binary classifier is trained for an individual horizon. To avoid leakage between highly correlated prefixes, trajectories are partitioned into training, validation, and test splits before prefix extraction. All prefixes derived from the same underlying trajectory therefore remain in the same split.

B.2 Training Protocol

RISA is trained offline in three sequential stages. Complete pseudocode for all three stages is provided in Appendix A.

Stage 1: multi-horizon risk prediction. The shared predictor g_ϕ is trained on trajectory prefixes using the categorical first-violation objective in Eq. (14). Given a prefix H_t, the predictor estimates the first-violation distribution $\boldsymbol{p}_t$. The horizon-specific risks $\widehat{R}^{(k)}_t$ are then obtained by cumulative summation according to Eq. (10).

Algorithm 3 Training the Risk-Informed Intervention Policy

Require: Training prefixes $\mathcal{P}$ with targets $\widetilde{d}_t$; fixed π_θ, g_ϕ, and $d_{\text{safe}}^{(\ell)}$; horizons $\mathcal{K}$ with weights $\{w_k\}$; policy π_ψ^{int}; group size G; reward weights $\lambda_r, \lambda_e, \lambda_s, \lambda_u, \lambda_c$; $\epsilon_{\text{clip}}, \epsilon$; updates E_{int}

Ensure: Trained intervention policy π_ψ^{int}

$k_M \leftarrow \max \mathcal{K}$; initialize ψ; $\psi_{\text{old}} \leftarrow \psi$

for $e = 1, \ldots, E_{\text{int}}$ **do**

  Sample $(H_t, \widetilde{d}_t) \in \mathcal{P}$ and compute $\boldsymbol{p}_t \leftarrow g_\phi(H_t)$, $\widehat{R}_t^{(k)} \leftarrow \sum_{j=1}^{k} p_t^{(j)}$ for all $k \in \mathcal{K}$

  Form $\widehat{\boldsymbol{R}}_t = [\widehat{R}_t^{(k_1)}, \ldots, \widehat{R}_t^{(k_M)}]$

  $\eta_t \leftarrow \widetilde{d}_t / k_M$ if $\widetilde{d}_t \in \{1, \ldots, k_M\}$, else $\eta_t \leftarrow 0$

  Generate matched no-intervention continuation $H_{t+1}^{(0)}$ using $\alpha_t^{(0)} = 0$ and compute $\widehat{R}_{t+1}^{(k,0)}$ for all $k \in \mathcal{K}$

  Sample $\alpha_t^{(1)}, \ldots, \alpha_t^{(G)} \sim \pi_{\psi_{\text{old}}}^{\text{int}}(\cdot \mid \widehat{\boldsymbol{R}}_t)$, $\alpha_t^{(i)} \in [0, 1]$

  for $i = 1, \ldots, G$ **do**

    Apply $\widetilde{h}_t^{(\ell,i)} \leftarrow h_t^{(\ell)} + \alpha_t^{(i)} d_{\text{safe}}^{(\ell)}$ and generate $H_{t+1}^{(i)}$

    Compute $\boldsymbol{p}_{t+1}^{(i)} \leftarrow g_\phi(H_{t+1}^{(i)})$ and $\widehat{R}_{t+1}^{(k,i)} \leftarrow \sum_{j=1}^{k} p_{t+1}^{(j,i)}$ for all $k \in \mathcal{K}$

    $$\Delta_t^{(i)} \leftarrow \sum_{k \in \mathcal{K}} w_k \left(\widehat{R}_{t+1}^{(k,0)} - \widehat{R}_{t+1}^{(k,i)} \right)$$

    Roll out candidate i and obtain $r_{\text{safe}}^{(i)}$ and $r_{\text{task}}^{(i)}$

    $q_t^{(i)} \leftarrow \lambda_r(1 + \lambda_e \eta_t)\Delta_t^{(i)} + \lambda_s r_{\text{safe}}^{(i)} + \lambda_u r_{\text{task}}^{(i)} - \lambda_c(\alpha_t^{(i)})^2$

  end for

  Compute group statistics $\overline{q}_t = \frac{1}{G}\sum_i q_t^{(i)}$ and $\sigma_t = \sqrt{\frac{1}{G}\sum_i (q_t^{(i)} - \overline{q}_t)^2}$

  for $i = 1, \ldots, G$ **do**

    $$A_t^{(i)} \leftarrow \frac{q_t^{(i)} - \overline{q}_t}{\sigma_t + \epsilon}, \qquad \rho_t^{(i)}(\psi) \leftarrow \frac{\pi_\psi^{\text{int}}(\alpha_t^{(i)} \mid \widehat{\boldsymbol{R}}_t)}{\pi_{\psi_{\text{old}}}^{\text{int}}(\alpha_t^{(i)} \mid \widehat{\boldsymbol{R}}_t)}$$

  end for

  $$\mathcal{L}_{\text{int}} \leftarrow -\frac{1}{G}\sum_{i=1}^{G} \min\Big(\rho_t^{(i)} A_t^{(i)}, \text{clip}(\rho_t^{(i)}, 1 - \epsilon_{\text{clip}}, 1 + \epsilon_{\text{clip}}) A_t^{(i)} \Big)$$

  Update ψ by minimizing $\mathcal{L}_{\text{int}}$; $\psi_{\text{old}} \leftarrow \psi$

end for

return π_ψ^{int}

Stage 2: safety-direction construction. Using only prefixes from the training split, we extract hidden representations from the acting model at intervention layer ℓ. Prefixes are partitioned into $\mathcal{P}_{\text{risk}}$ and $\mathcal{P}_{\text{safe}}$ according to $\widetilde{d}_t$, and the normalized safety direction $d_{\text{safe}}^{(\ell)}$ is computed using Eq. (19). The resulting direction remains fixed during subsequent intervention-policy training and inference.

Stage 3: adaptive intervention-policy training. The acting model π_θ, risk predictor g_ϕ, and safety direction $d_{\text{safe}}^{(\ell)}$ are frozen. The intervention policy π_ψ^{int} is optimized using the trajectory-aware reward in Eq. (24) and the group-relative objective in Eq. (27).

For each training prefix, G intervention strengths are sampled from the intervention policy and evaluated relative to a separate matched no-intervention continuation with $\alpha_t^{(0)} = 0$. The no-intervention continuation serves only as the reference used to measure predicted risk reduction and is not included in the G samples used for group-relative normalization. Future safety outcomes, first-violation offsets, and rollout outcomes are used only to construct the offline training reward. At inference time, the policy receives only the predicted multi-horizon risk vector $\widehat{\boldsymbol{R}}_t$ computed from the observed prefix H_t.

B.3 Backbone-Specific Training

RISA is trained independently for each acting-model backbone. For Qwen3-4B, Qwen2.5-14B, Gemma-2-9B, and LLaMA-8B, we separately train the risk predictor g_ϕ, select the intervention layer ℓ, construct $d_{\text{safe}}^{(\ell)}$, and optimize the intervention policy π_ψ^{int}.

No risk-predictor, representation-direction, or intervention-policy parameters are transferred across model families. This avoids assuming that either trajectory-risk representations or safety-relevant hidden geometry are directly shared across different backbone architectures.

B.4 Baseline and Ablation Configuration

We use the same baseline set described in Section 4.1.

Base Agent. The acting model is evaluated without an additional runtime safeguard.

ToolSafe. A step-level safeguard that evaluates proposed tool use using the available interaction context (Mou et al., 2026).

TRACE. A trajectory-aware safeguard that aggregates safety-relevant evidence across the observed interaction history (Hong et al., 2026).

TRACES. A proactive trajectory-level safeguard that estimates future risk from partial interaction histories (Li et al., 2026a).

Safety Rubric. A rubric-based agent safeguard that evaluates agent behavior against explicit safety criteria (Loye et al., 2026).

We additionally evaluate the following controlled variants of RISA.

Fixed Steering. The learned direction $d_{\text{safe}}^{(\ell)}$ is applied using the same non-zero steering strength at every decision step.

Threshold Steering. A fixed steering strength is activated when a scalar predicted-risk value exceeds a threshold selected on the validation split.

Single-Horizon RISA. The intervention policy is trained using only the longest-horizon prediction $\widehat{R}_t^{(k_M)}$, rather than the full vector $\widehat{\boldsymbol{R}}_t$.

RISA without lead-time reward. This variant removes the early-intervention weighting term by setting $\lambda_e = 0$ in Eq. (24), while leaving the remaining training procedure unchanged.

B.5 Evaluation Metrics

Multi-horizon risk forecasting. For each horizon $k \in \mathcal{K}$, we evaluate $\widehat{R}_t^{(k)}$ against the corresponding binary target. We report area under the receiver operating characteristic curve (**AUROC**) and area under the precision–recall curve (**AUPRC**) separately for each prediction horizon.

Because g_ϕ directly models the categorical first-violation distribution, we additionally report negative log-likelihood (**NLL**). Conditioned on a violation occurring within the maximum forecasting window, the expected first-violation offset is

$$\widehat{d}_t = \frac{\sum_{j=1}^{k_M} j\, p_t^{(j)}}{1 - p_t^{(\varnothing)}}, \tag{29}$$

and we report mean absolute error (**MAE**) between $\widehat{d}_t$ and d_t over test prefixes whose first violation occurs within the forecasting window.

End-to-end safety and utility. End-to-end metrics are computed over completed evaluation trajectories. Let $\mathcal{T}_{\text{unsafe}}$ denote the safety-challenging evaluation set and $\mathcal{T}_{\text{benign}}$ the benign evaluation set. Unsafe Completion Rate (**UCR**) is measured on the safety-challenging set:

$$\text{UCR} = \frac{1}{|\mathcal{T}_{\text{unsafe}}|} \sum_{H_T \in \mathcal{T}_{\text{unsafe}}} \mathbb{I}\left[H_T \text{ results in an unsafe completion}\right]. \tag{30}$$

Safe Completion Rate (**SCR**) and Over-Refusal Rate (**ORR**) are measured on benign tasks:

$$\text{SCR} = \frac{1}{|\mathcal{T}_{\text{benign}}|} \sum_{H_T \in \mathcal{T}_{\text{benign}}} \mathbb{I}\left[H_T \text{ results in a successful safe completion}\right], \tag{31}$$

$$\text{ORR} = \frac{1}{|\mathcal{T}_{\text{benign}}|} \sum_{H_T \in \mathcal{T}_{\text{benign}}} \mathbb{I}\left[H_T \text{ is unnecessarily refused}\right]. \tag{32}$$

Task utility is computed using the task-success evaluation provided by the corresponding benchmark. Because UCR is evaluated on safety-challenging trajectories while SCR and ORR are evaluated on benign trajectories, these quantities are not required to sum to one. Together, they character-

Table 4: Safety and utility across long-horizon agent-safety benchmarks for LLaMA-8B and Qwen2.5-14B. Base Agent provides the unsafeguarded utility reference. Best safeguarded result for each model and metric is shown in bold.

		LLaMA-8B				Qwen2.5-14B			
Benchmark	**Method**	**UCR↓**	**Utility↑**	**ORR↓**	**SCR↑**	**UCR↓**	**Utility↑**	**ORR↓**	**SCR↑**
BLINDSPOT	Base Agent	0.401	0.861	0.049	0.548	0.387	0.879	0.044	0.566
	ToolSafe	0.342	0.758	0.142	0.659	0.321	0.779	0.131	0.676
	TRACE	0.348	0.776	0.125	0.648	0.274	0.798	0.116	0.711
	TRACES	0.249	0.789	0.111	0.716	0.232	0.762	0.102	0.733
	SafetyRubric	0.226	0.771	0.129	0.691	0.285	0.796	0.119	0.704
	RISA	**0.106**	**0.820**	**0.074**	**0.794**	**0.098**	**0.822**	**0.068**	**0.812**
MT-AgentRisk	Base Agent	0.372	0.873	0.046	0.579	0.354	0.891	0.041	0.597
	ToolSafe	0.243	0.807	0.116	0.674	0.226	0.829	0.105	0.691
	TRACE	0.207	0.823	0.096	0.704	0.248	0.845	0.088	0.681
	TRACES	0.176	0.816	0.103	0.678	0.161	0.787	0.095	0.744
	SafetyRubric	0.214	0.795	0.121	0.718	0.139	**0.817**	0.111	0.711
	RISA	**0.096**	**0.844**	**0.068**	**0.752**	**0.087**	0.815	**0.061**	**0.819**
LongDS-Bench	Base Agent	0.338	0.866	0.044	0.611	0.322	0.884	0.039	0.628
	ToolSafe	0.219	0.801	0.109	0.694	0.203	0.823	0.099	0.710
	TRACE	0.255	0.817	0.092	0.669	0.173	0.831	0.083	0.735
	TRACES	0.162	0.809	0.100	0.740	0.148	0.781	0.091	0.756
	SafetyRubric	0.139	0.788	0.119	0.708	0.127	0.811	0.096	0.703
	RISA	**0.089**	**0.829**	**0.064**	**0.811**	**0.080**	**0.848**	**0.058**	**0.827**

ize the safety–utility trade-off without conflating unsafe-task robustness with benign-task preservation.

Warning lead time. For evaluation trajectories whose matched no-intervention continuation reaches a safety violation, warning lead time is defined as

$$\Delta_{\text{lead}} = t^{(0)}_{\text{unsafe}} - t_{\text{int}}, \tag{33}$$

where $t^{(0)}_{\text{unsafe}}$ is the first violating step in the matched no-intervention continuation and t_{int} is the first decision step at which RISA applies a non-zero intervention, $\alpha_t > 0$. Larger values indicate that the safeguard intervenes earlier relative to the impending unsafe event.

Runtime overhead. Runtime overhead is measured as the additional wall-clock latency per agent decision relative to the Base Agent. We additionally report the number of safeguard-related model or predictor evaluations required per decision, separating safety-monitoring cost from the acting model's normal generation cost.

C Additional Results Across Model Families

Section 4.3 reports the results for Qwen3-4B and Gemma-2-9B. We further evaluate LLaMA-8B and Qwen2.5-14B to examine whether the observed safety–utility trade-off extends across model families and scales. For each backbone, the risk predictor, safety direction, and intervention policy are trained independently, following Appendix B.3.

Across the additional backbones, the results exhibit the same overall safety-utility pattern observed for Qwen3-4B and Gemma-2-9B. RISA achieves the lowest unsafe-completion and over-refusal rates across all evaluated settings while retaining competitive safeguarded task utility. On LLaMA-8B, RISA obtains a UCR of 0.106 on BLINDSPOT while retaining utility of 0.820. On MT-AgentRisk and LongDS-Bench, the corresponding UCRs are 0.096 and 0.089, with utilities of 0.844 and 0.829, respectively.

The same qualitative behavior is observed for Qwen2.5-14B. RISA achieves UCRs of 0.098, 0.087, and 0.080 on BLINDSPOT, MT-AgentRisk, and LongDS-Bench, respectively, while maintaining utilities of 0.822, 0.815, and 0.848. On BLINDSPOT and LongDS-Bench, RISA also achieves the highest safeguarded utility, whereas on MT-AgentRisk its utility remains comparable to the strongest baseline while substantially reducing unsafe completion and over-refusal.

Table 5: First-violation forecasting performance across long-horizon agent-safety benchmarks. MAE is measured in interaction steps and is computed only on prefixes whose first violation occurs within the maximum forecasting horizon k_M.

Benchmark	**NLL↓**	**MAE↓**
BLINDSPOT	0.842	1.28
MT-AgentRisk	0.913	1.41
LongDS-Bench	0.978	1.56
Average	0.911	1.42

Table 6: Calibration of cumulative multi-horizon risk predictions on held-out trajectory prefixes.

Horizon	**ECE↓**	**Brier↓**
$k = 1$	0.041	0.112
$k = 2$	0.048	0.128
$k = 4$	0.057	0.146
$k = 8$	0.069	0.171

These results indicate that RISA's gains are not restricted to a particular model architecture or parameter scale. Across both additional backbones, adaptive multi-horizon intervention consistently reduces unsafe completion and over-refusal while preserving strong task utility. The results therefore support the broader applicability of RISA across different model families and scales.

D Additional Risk Predictor Analysis

This section provides additional analysis of the shared risk predictor g_ϕ beyond the horizon-wise AUROC and AUPRC results reported in Section 4.2. We examine the accuracy of the underlying first-violation distribution and the calibration of the cumulative multi-horizon risk estimates.

D.1 First-Violation Forecasting

The shared predictor g_ϕ models the categorical distribution over the offset to the first future safety violation, as defined in Section 3.1. While the main evaluation measures the cumulative risks $\widehat{R}_t^{(k)}$ at each prediction horizon, we additionally evaluate the underlying first-violation distribution directly. Negative log-likelihood (**NLL**) evaluates the complete categorical first-violation distribution. To measure temporal localization, we use the conditional expected first-violation offset defined in Eq. (29) and report mean absolute error (**MAE**) against the observed first-violation offset d_t. The final row reports the macro-average across the three benchmarks.

The predictor localizes the first upcoming violation with an average error of 1.42 interaction steps across the evaluated benchmarks. BLINDSPOT yields the lowest temporal prediction error, while forecasting is more difficult on MT-AgentRisk and LongDS-Bench.

The quantity $\widehat{d}_t$ is used only to evaluate the temporal information captured by the learned first-violation distribution. It is not provided to the deployed intervention policy. At inference time, the policy operates only on the cumulative multi-horizon risk vector $\widehat{\boldsymbol{R}}_t$ derived from the observed trajectory prefix.

D.2 Calibration of Multi-Horizon Risk

Because the intervention policy uses predicted risk probabilities to select the intervention strength, we additionally evaluate calibration of the cumulative risk estimates. For each horizon $k \in \mathcal{K}$, we compare $\widehat{R}_t^{(k)}$ with the binary event

$$\mathbb{I}\left[\widetilde{d}_t \in \{1, \ldots, k\}\right],$$

which indicates whether the first future violation occurs within horizon k. We report Expected Calibration Error (**ECE**) and Brier score for each prediction horizon.

Calibration error increases gradually with the prediction horizon: ECE rises from 0.041 at $k = 1$ to 0.069 at $k = 8$, while the Brier score increases from 0.112 to 0.171. This trend is consistent with the horizon-wise forecasting results in Section 4.2, where predicting more distant safety events

Table 7: Layer-wise separability of safe and future-risk prefixes for Qwen3-4B on BLINDSPOT. Relative depth is normalized by the total number of transformer layers. AUROC measures discrimination between the two prefix groups using the projection score $s_t^{(\ell)}$.

Relative Depth	Safe Proj.↑	Risk Proj.↓	AUROC↑
25%	0.21	0.08	0.612
50%	0.47	−0.19	0.741
75%	0.83	−0.51	0.861
100%	0.71	−0.39	0.824

Table 8: Matched-magnitude steering controls for Qwen3-4B on BLINDSPOT. All non-zero perturbations use the same steering magnitude.

Direction	UCR↓	Utility↑	ORR↓
No Steering	0.432	0.842	0.057
Random Direction	0.407	0.816	0.073
$-d_{\text{safe}}^{(\ell)}$	0.468	0.823	0.061
$+d_{\text{safe}}^{(\ell)}$	0.276	0.817	0.089

is also more difficult. Nevertheless, the gradual degradation indicates that the shared first-violation predictor retains useful probabilistic information across all evaluated horizons.

E Safety-Direction Analysis

We analyze whether the learned direction $d_{\text{safe}}^{(\ell)}$ captures safety-relevant structure in the acting model's hidden representation space and whether steering along this direction produces behaviorally meaningful changes. Unless otherwise stated, the analyses in this section use Qwen3-4B on BLINDSPOT.

E.1 Layer-Wise Separability

For a held-out trajectory prefix H_t, we measure its projection onto the safety direction at layer ℓ as

$$s_t^{(\ell)} = \left\langle h_t^{(\ell)}, d_{\text{safe}}^{(\ell)} \right\rangle. \tag{34}$$

For every candidate layer, the direction $d_{\text{safe}}^{(\ell)}$ is constructed using training prefixes only, while separability is evaluated on held-out prefixes. Higher projection scores indicate greater alignment with the safe representation centroid.

As shown in Table 7, separability improves substantially in later layers, reaching its highest value at approximately 75% of model depth rather than at the final transformer layer. This supports selecting the intervention layer using validation data instead of assuming that the final layer provides the most safety-relevant representation.

E.2 Matched-Magnitude Direction Control

Representation separability alone does not establish that steering along the learned direction meaningfully changes agent behavior. We therefore compare $+d_{\text{safe}}^{(\ell)}$ with no steering, a norm-matched random direction, and the reversed direction $-d_{\text{safe}}^{(\ell)}$, while keeping the perturbation magnitude fixed across steering conditions.

Table 8 shows that positive steering along $d_{\text{safe}}^{(\ell)}$ reduces UCR from 0.432 to 0.276, compared with 0.407 for a norm-matched random direction. Reversing the learned direction instead increases UCR to 0.468. This directional asymmetry indicates that the observed safety improvement is associated with the orientation of the learned representation direction rather than merely the magnitude of a generic hidden-state perturbation.

The same table also shows that the reduction in unsafe completion is accompanied by a moderate decrease in task utility and an increase in over-refusal relative to no steering. This further motivates the adaptive intervention policy, which controls when and how strongly the safety direction is applied rather than steering every prefix uniformly.

Table 9: Mean intervention strength for held-out Qwen3-4B BLINDSPOT prefixes with similar eight-step risk but different one-step risk.

$\widehat{R}_t^{(1)}$	$\widehat{R}_t^{(8)}$	**Mean** α_t
0.12	0.78	0.31
0.34	0.79	0.47
0.61	0.80	0.68
0.79	0.81	0.76

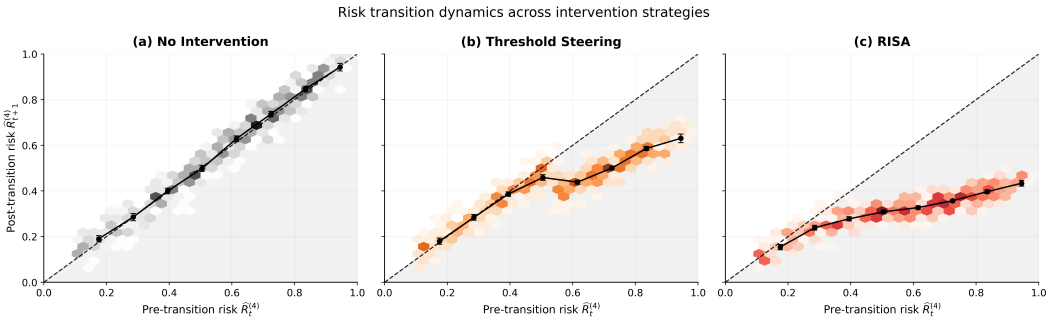

Figure 3: **Matched risk-transition dynamics.** Predicted horizon-4 risk before and after one transition under **(a)** no intervention, **(b)** threshold steering, and **(c)** RISA. The dashed diagonal denotes no change in predicted risk; points below the diagonal indicate lower predicted future risk after the transition. Hexagonal bins indicate trajectory-prefix density, and the black curve shows the binned mean post-transition risk as a function of pre-transition risk.

E.3 Does the Intervention Policy Use the Full Risk Profile?

The intervention policy receives the complete multi-horizon risk vector

$$\widehat{\boldsymbol{R}}_t = \left[\widehat{R}_t^{(1)}, \widehat{R}_t^{(2)}, \widehat{R}_t^{(4)}, \widehat{R}_t^{(8)}\right].$$

To examine whether the learned policy makes use of information beyond the longest-horizon risk alone, we compare held-out prefixes with similar $\widehat{R}_t^{(8)}$ values but substantially different near-term risk $\widehat{R}_t^{(1)}$.

As shown in Table 9, the eight-step risk remains nearly constant while the mean intervention strength increases from 0.31 to 0.76 as one-step risk increases from 0.12 to 0.79. Thus, prefixes with comparable long-horizon risk can induce substantially different intervention strengths depending on their near-term risk profile.

Together with the single-horizon ablation in Section 4.4, Table 9 provides evidence that the learned intervention policy does not operate solely as a threshold on $\widehat{R}_t^{(k_M)}$. Instead, its decisions vary with the temporal distribution of predicted risk represented by $\widehat{\boldsymbol{R}}_t$.

F Matched Risk-Transition Analysis

To examine how intervention affects short-term trajectory evolution, we perform a matched-prefix transition analysis on held-out prefixes. Starting from the same prefix H_t, we generate one subsequent transition under three conditions: no intervention, threshold steering, and RISA. All resulting prefixes are evaluated using the same frozen risk predictor g_ϕ, allowing changes in predicted risk to be compared from a common initial state.

For each condition, we compare the horizon-4 risk before intervention, $\widehat{R}_t^{(4)}$, with the risk after the resulting transition, $\widehat{R}_{t+1}^{(4)}$.

As shown in Figure 3, the no-intervention condition remains comparatively close to the diagonal, indicating that predicted risk often persists across the next transition. Threshold steering produces a sharper change around its activation region, consistent with its discrete triggering rule. In contrast, RISA produces a smoother downward shift across medium- and high-risk prefixes, reflecting the adaptive variation of intervention strength with the multi-horizon risk profile.

Table 10: Ablation of intervention-reward components for LLaMA-8B on BLINDSPOT. Each variant removes a single reward component while keeping the remaining training procedure fixed.

Reward Variant	UCR↓	Utility↑	ORR↓
w/o Risk Reduction	0.161	0.807	0.089
w/o Task Utility	0.091	0.716	0.198
w/o Intervention Cost	0.087	0.739	0.174
Full RISA	0.106	**0.820**	**0.074**

The contrast is especially informative for prefixes with comparable initial horizon-4 risk. Rather than applying the same modification to all prefixes above a fixed threshold, RISA can assign different intervention strengths based on the full risk vector $\widehat{\boldsymbol{R}}_t$. The resulting transition dynamics are therefore consistent with the policy analysis in Section E.3, where intervention strength varies even when long-horizon risk remains similar.

Because the same frozen predictor is used to score both the pre-transition and post-transition prefixes, Figure 3 measures changes in *predicted* future risk rather than independently adjudicated downstream safety outcomes. We therefore treat this analysis as a mechanistic diagnostic of how intervention changes the local trajectory state and interpret it together with the end-to-end safety and utility results in Section 4.3.

G Intervention Reward Ablations

The intervention reward in Eq. (24) combines predicted risk reduction, safety outcome, task utility, early-intervention credit, and intervention cost. To examine the role of the individual terms, we remove one component at a time while leaving the remaining intervention-policy training procedure unchanged. Unless otherwise stated, these ablations use LLaMA-8B on BLINDSPOT, matching the intervention analysis in Section 4.4.

As shown in Table 10, removing the risk-reduction term increases UCR from 0.106 to 0.161, indicating that directly rewarding reductions in predicted future risk provides an important training signal for safer intervention decisions.

Removing the task-utility term produces a lower raw UCR of 0.091, but utility drops substantially from 0.820 to 0.716 and ORR rises from 0.074 to 0.198. A similar pattern appears when the intervention-cost term is removed: UCR decreases to 0.087, while utility falls to 0.739 and ORR increases to 0.174. These results show that optimizing unsafe completion alone is insufficient, since the policy can obtain lower UCR by intervening more aggressively at the expense of useful behavior.

The complete reward therefore balances complementary objectives: predicted risk reduction encourages transitions toward safer future states, while task utility and intervention cost discourage unnecessarily strong or indiscriminate steering. This trade-off is consistent with the main end-to-end evaluation, where RISA is designed to improve safety without relying on excessive refusal.

G.1 Effect of Early-Intervention Credit

The early-intervention component is analyzed separately because its primary effect is on *when* the policy intervenes rather than only on whether the final trajectory is safe. As shown in Table 3, setting $\lambda_e = 0$ reduces mean warning lead time from 3.29 to 1.94 interaction steps. The changes in UCR, utility, and ORR are comparatively smaller.

This behavior supports the intended role of the lead-time factor in Eq. (24): it assigns additional credit to successful risk reduction that occurs earlier in the trajectory, rather than simply encouraging larger intervention strengths. The result complements the matched risk-transition analysis in Section F, which examines how intervention changes predicted risk locally after a single transition.

H Additional Intervention-Policy Analysis

We further analyze the intervention-policy optimization by varying the GRPO candidate-group size and the relative weighting of the reward components. These experiments isolate the sensitivity of policy learning while keeping the acting model, risk predictor, safety direction, and remaining training configuration fixed.

Table 11: Sensitivity to GRPO candidate-group size G. Training cost is reported relative to the $G = 2$ configuration.

G	UCR↓	Utility↑	ORR↓	Training Cost×
2	0.154	0.804	0.096	1.0
4	0.132	0.808	0.089	1.8
8	0.121	0.811	0.083	3.4
16	0.118	0.810	0.084	6.6

Table 12: Sensitivity to intervention-policy reward weighting. The balanced configuration corresponds to the reward weighting used for the reference policy in this analysis.

Configuration	UCR↓	Utility↑	ORR↓
Safety-heavy	0.104	0.748	0.173
Utility-heavy	0.166	0.834	0.066
Early-heavy	0.119	0.792	0.101
Cost-heavy	0.151	0.826	0.071
Balanced (RISA)	0.121	0.811	0.083

H.1 GRPO Group Size

For each training prefix, GRPO compares G candidate intervention strengths sampled from the current intervention policy. We vary G to measure the trade-off between the number of within-prefix intervention comparisons and the rollout cost required for policy optimization.

As shown in Table 11, increasing the candidate-group size from $G = 2$ to $G = 8$ progressively reduces UCR from 0.154 to 0.121 and ORR from 0.096 to 0.083, while maintaining similar task utility. Increasing the group further to $G = 16$ yields only a small additional UCR reduction, from 0.121 to 0.118, while nearly doubling the training cost relative to $G = 8$.

These results indicate diminishing returns from additional candidate rollouts. We therefore use $G = 8$ in the main intervention-policy training configuration as a balance between policy performance and training cost.

H.2 Reward-Weight Sensitivity

We next vary the relative weighting of the terms in Eq. (24) around the configuration used for the main policy training. The GRPO group size is fixed to $G = 8$ for this analysis. Each variant increases the relative emphasis of one reward component while leaving the remaining policy-training procedure unchanged.

Table 12 exposes the expected trade-off among the reward components. Increasing the relative emphasis on safety reduces UCR to 0.104, but utility decreases to 0.748 and ORR increases to 0.173, indicating substantially more conservative intervention. Conversely, the utility-heavy configuration achieves the highest task utility, 0.834, and the lowest ORR, 0.066, but increases UCR to 0.166.

Increasing the early-intervention weight produces a smaller change in UCR but reduces utility and increases ORR relative to the balanced configuration. A stronger intervention-cost penalty has the opposite effect: it preserves utility and limits over-refusal, but increases UCR to 0.151 by discouraging stronger intervention.

Overall, the sensitivity results show that no single reward component simultaneously optimizes safety, utility, and intervention conservativeness. The balanced configuration provides an intermediate operating point and is used as the reference policy for these sensitivity analyses. Together with the component-removal results in Section G, these experiments illustrate the distinct roles of the reward terms in shaping intervention behavior.

I Implementation Details

This section provides implementation details for the trajectory preprocessing, multi-horizon risk predictor, representation steering, and intervention-policy components of RISA. Unless otherwise stated, hyperparameters are selected using the validation split and fixed before test evaluation. The same configuration is used across benchmarks for a given backbone unless explicitly stated otherwise.

Table 13: Multi-horizon risk-predictor training configuration.

Parameter	Value
Prediction horizons $\mathcal{K}$	$\{1, 2, 4, 8\}$
Maximum horizon k_M	8
Output classes	9
Predictor depth	2 MLP layers
Hidden dimension	256
Activation	GELU
Dropout	0.1
Training objective	Categorical NLL
Optimizer	AdamW
Learning rate	1×10^{-4}
Batch size	256 prefixes
Weight decay	1×10^{-2}
Maximum epochs	20
Gradient clipping	1.0
Early-stopping patience	3 epochs
Selection criterion	Validation mean AUPRC

I.1 Trajectory Preprocessing

Each completed trajectory H_T is decomposed into observed prefixes H_t with $t \in \{1, \ldots, T-1\}$. We exclude H_T because no future decision step remains from which a first-violation target can be defined. For every retained prefix, we construct the categorical first-violation target $\widetilde{d}_t$ as described in Eq. (13).

We use

$$\mathcal{K} = \{1, 2, 4, 8\}, \qquad k_M = \max \mathcal{K} = 8.$$

If the first future violation occurs at offset $j \leq k_M$, then $\widetilde{d}_t = j$. Otherwise, $\widetilde{d}_t = \varnothing$, including when no violation occurs before the trajectory terminates.

For horizon-wise evaluation, the binary event associated with horizon k is derived directly as

$$\mathbb{I}\left[\widetilde{d}_t \in \{1, \ldots, k\}\right].$$

No independent horizon-specific classifiers are trained.

Trajectories are partitioned into training, validation, and test splits before prefix extraction. Consequently, prefixes originating from the same underlying trajectory never appear in different partitions.

I.2 Multi-Horizon Risk Predictor

Given an observed prefix H_t, the shared predictor g_ϕ estimates the categorical first-violation distribution

$$\boldsymbol{p}_t = \left[p_t^{(1)}, \ldots, p_t^{(k_M)}, p_t^{(\varnothing)}\right].$$

The corresponding cumulative risk for any $k \in \mathcal{K}$ is

$$\widehat{R}_t^{(k)} = \sum_{j=1}^{k} p_t^{(j)}.$$

Because all horizon-specific risks are derived from the same categorical distribution, monotonicity with respect to prediction horizon follows by construction and no additional monotonicity regularizer is required.

Predictor architecture. For each acting-model backbone, we use its final-token hidden representation of H_t as the predictor input. Gradients from the risk-prediction objective are not propagated into the acting model. The representation is passed through a two-layer MLP with hidden dimension 256, GELU activation, and dropout 0.1. A linear output layer produces $k_M + 1 = 9$ logits followed by a softmax over the eight possible first-violation offsets and the no-violation outcome $\varnothing$.

Optimization. The predictor is optimized using the categorical negative log-likelihood in Eq. (14). Hyperparameters are listed in Table 13.

Table 14: Representation-steering configuration.

Parameter	Value
Representation	Final token of H_t
Candidate depths	$25\%, 50\%, 75\%, 100\%$
Layer selection metric	Validation projection AUROC
Direction estimator	Difference of group means
Direction normalization	ℓ_2 normalization
Direction estimation split	Training only
Intervention-strength space	$[0, 1]$
Direction update during policy training	None

The validation criterion is the mean AUPRC over the four forecasting horizons. The checkpoint with the highest validation value is retained for all downstream direction construction and intervention-policy training.

I.3 Safety-Direction Construction

For each acting-model backbone, we extract the final-token hidden state $h_t^{(\ell)}$ for every training prefix at a set of candidate transformer layers. Prefixes are partitioned into

$$\mathcal{P}_{\text{risk}} = \left\{ t : \widetilde{d}_t \in \{1, \dots, k_M\} \right\}, \qquad \mathcal{P}_{\text{safe}} = \left\{ t : \widetilde{d}_t = \varnothing \right\}.$$

The safety direction $d_{\text{safe}}^{(\ell)}$ is computed using Eq. (19). Group centroids are estimated using training prefixes only; validation and test representations are never used to estimate the direction.

We evaluate candidate layers at approximately 25%, 50%, 75%, and 100% of transformer depth. The intervention layer is selected independently for each backbone according to validation-set safe-versus-risk separability measured by projection AUROC. Once the layer is selected, both ℓ and $d_{\text{safe}}^{(\ell)}$ remain fixed during intervention-policy training and test-time evaluation.

I.4 Intervention Policy

At decision step t, the intervention policy receives only

$$\widehat{\boldsymbol{R}}_t = \left[\widehat{R}_t^{(1)}, \widehat{R}_t^{(2)}, \widehat{R}_t^{(4)}, \widehat{R}_t^{(8)} \right].$$

No hidden trajectory labels or future information are included in the policy input.

Policy architecture. We parameterize π_ψ^{int} as a lightweight two-layer MLP with hidden dimension 64 and GELU activations. Its output parameterizes a Beta distribution over the bounded continuous intervention strength

$$\alpha_t \in [0, 1].$$

Specifically, the policy produces two positive concentration parameters through a softplus transformation, and candidate strengths are sampled as

$$\alpha_t \sim \text{Beta}\left(a_\psi(\widehat{\boldsymbol{R}}_t), b_\psi(\widehat{\boldsymbol{R}}_t) \right).$$

A small constant 10^{-4} is added to both concentration parameters for numerical stability.

During evaluation, we use the mean of the learned Beta distribution,

$$\alpha_t = \frac{a_\psi(\widehat{\boldsymbol{R}}_t)}{a_\psi(\widehat{\boldsymbol{R}}_t) + b_\psi(\widehat{\boldsymbol{R}}_t)},$$

which provides deterministic intervention strengths while preserving the continuous action space used during training.

The deployable policy never observes $\widetilde{d}_t$, d_t, future trajectory length, future actions, future tool outputs, or future safety outcomes.

I.5 GRPO Training

For each sampled training prefix, we draw G candidate intervention strengths from

$$\pi_{\psi_{\text{old}}}^{\text{int}} \left(\cdot \mid \widehat{\boldsymbol{R}}_t \right).$$

Table 15: GRPO intervention-policy training configuration.

Parameter	Value
Policy input dimension	4
Policy depth	2 MLP layers
Hidden dimension	64
Policy distribution	Beta
Intervention range	$[0, 1]$
Group size G	8
Optimizer	AdamW
Policy learning rate	3×10^{-4}
Weight decay	1×10^{-2}
GRPO clipping ϵ_{clip}	0.2
Gradient clipping	1.0
Prefix batch size	64
Training updates	5,000
Policy epochs per sampled batch	1
λ_r	1.0
λ_e	0.5
λ_s	1.0
λ_u	1.0
λ_c	0.1
Numerical constant ϵ	10^{-6}

The matched no-intervention action $\alpha_t^{(0)} = 0$ is generated separately and serves only as the reference continuation for computing the predicted risk reduction in Eq. (22). It is not included among the G candidates used to compute the group-relative reward mean or standard deviation.

Candidate and no-intervention continuations use identical generation settings. When stochastic decoding is enabled, matched continuations use the same decoding seed so that differences are attributable as closely as possible to the intervention rather than sampling noise.

For each candidate, we compute the reward in Eq. (24). The risk-reduction term $\Delta_t^{(i)}$ lies in $[-1, 1]$, while the task and safety rewards are normalized to $[0, 1]$. The normalized lead-time factor η_t and intervention cost $(\alpha_t^{(i)})^2$ also lie in $[0, 1]$.

We use the reference reward weights

$$\lambda_r = 1.0, \qquad \lambda_e = 0.5, \qquad \lambda_s = 1.0, \qquad \lambda_u = 1.0, \qquad \lambda_c = 0.1.$$

Group-relative advantages are computed according to Eq. (26). The acting model π_θ, risk predictor g_ϕ, intervention layer ℓ, and safety direction $d_{\text{safe}}^{(\ell)}$ remain frozen throughout this stage; only the intervention-policy parameters ψ are updated.

We use $G = 8$ because the group-size analysis in Table 11 shows diminishing performance gains beyond eight candidate interventions relative to the additional rollout cost.

Privileged training information. Complete trajectories are used only to construct offline supervision and rewards. In particular, $\widetilde{d}_t$, d_t, future safety labels, and candidate rollout outcomes are never provided as inputs to the deployed intervention policy. At inference time, intervention depends exclusively on the observed prefix through $\widehat{\boldsymbol{R}}_t$.

I.6 Acting Models and Decoding

We evaluate Qwen3-4B, Gemma-2-9B, LLaMA-8B, and Qwen2.5-14B. The acting-model parameters remain frozen throughout risk-predictor training, safety-direction construction, and intervention-policy optimization. The Base Agent and all safeguarded variants use the same generation configuration for a given backbone and benchmark.

A trajectory terminates when the benchmark environment reaches a terminal state, the task is completed, a terminal safety outcome is reached, or the benchmark-specific maximum interaction length is exceeded. Tool outputs and environment modifications persist across steps, so each continuation is evaluated under the resulting stateful environment rather than as an independent prompt.

Table 16: Generation configuration used for agent rollouts.

Parameter	Value
Temperature	0.2
Top-p	0.95
Maximum new tokens per decision	256
Maximum trajectory steps	Benchmark-defined
Stop criteria	Benchmark-defined
Tool-call execution	Stateful
Matched-rollout decoding	Same seed and settings
Evaluation policy action	Beta mean
Precision	bfloat16

I.7 Reproducibility

All hyperparameter selection, intervention-layer selection, and early stopping decisions are performed using the validation split. Test data are used only for final evaluation. Unless otherwise stated, reported results are averaged over three independent random seeds. The same trajectory splits and evaluation instances are used for all methods within a given backbone-benchmark pair.

For matched intervention comparisons, all methods start from the same trajectory prefix and use the same environment state and generation configuration. This pairing reduces variance due to differences in initial conditions and isolates the effect of the intervention method.

J Related Work

Safety in tool-using and long-horizon agents. The transition from conversational LLMs to agents that invoke tools and modify external environments expands the safety problem from harmful text generation to potentially consequential sequences of actions. Existing benchmarks have studied harmful tool use, prompt injection, unsafe action execution, and other agent-specific failure modes (Andriushchenko et al., 2025; Debenedetti et al., 2024). More recent work shows that these risks become particularly challenging in multi-turn settings because harmful objectives can be decomposed across individually innocuous interactions. MT-AgentRisk, for example, systematically transforms harmful tasks into multi-turn tool-use sequences and observes substantial degradation in agent safety relative to single-turn settings (Li et al., 2026b). These findings motivate reasoning about the trajectory as a whole rather than judging each interaction independently.

Long-horizon safety further introduces delayed and compositional failure patterns. An early action may be locally permissible yet change the environment or accumulated context in a way that makes a later violation increasingly likely. TRACE characterizes this setting in terms of sparse, delayed, and compositional safety evidence and learns to compress trajectory-level evidence for subsequent safety judgment (Hong et al., 2026). Such results establish that safety-relevant information can be distributed over long interaction histories. RISA builds on this observation but addresses a different question: rather than only determining whether an observed trajectory is unsafe, we seek to estimate how likely a violation is to occur over multiple future horizons from the currently observed prefix.

Runtime guardrails and trajectory-aware safety monitoring. A large body of work develops runtime safeguards that monitor agent behavior without modifying the underlying task policy. General-purpose guardrails such as Llama Guard and ShieldGemma primarily classify the safety of observed inputs or outputs (Inan et al., 2023; Zeng et al., 2024), while agent-specific systems extend monitoring to plans, tool calls, and execution trajectories (Xiang et al.; Chen et al., 2025; Chennabasappa et al., 2025). ToolSafe moves this protection closer to the point of execution: its TS-Guard reasons over interaction history and evaluates a proposed tool invocation before the action is executed, while TS-Flow uses the resulting feedback to alter subsequent agent reasoning (Mou et al., 2026). This form of step-level guarding is effective when the pending action already exposes sufficient evidence of risk.

Trajectory-aware approaches instead aggregate evidence over longer contexts. TRACE learns a latent trajectory representation that emphasizes risk-relevant segments while retaining access to the original trajectory (Hong et al., 2026). RUBAS takes a complementary training-time perspective, using structured trajectory-level rubrics covering tool-use safety, argument safety, response safety, and helpfulness as reinforcement-learning rewards (Loye et al., 2026). Together, these approaches improve reasoning about observed agent behavior, but they do not directly provide the combination of multi-timescale future-risk estimates and risk-conditioned graded intervention used by RISA.

Predictive safeguards and early risk forecasting. Recent work has begun to shift agent safeguards from retrospective detection toward prediction from partial trajectories. AgentForesight formulates online auditing for multi-agent systems, where an auditor observes only the current prefix and attempts to raise an alarm at the earliest decisive error (Zhang et al., 2026). Its objective is primarily early localization of trajectory failures rather than forecasting safety violations at several future horizons, but it demonstrates the value of acting before errors propagate through a long execution.

More directly, Recast formulates multi-turn safety as trajectory-level forecasting rather than current-state classification (Lin et al., 2026a). It retrieves short- and long-term risk evidence, models temporal risk evolution, and predicts a distribution over future risk-emergence turns. JANUS similarly argues for predictive rather than reactive guarding: from a partial trajectory, it anticipates safety-relevant future continuations and jointly uses the observed prefix and anticipated future to adjudicate whether an action should be blocked (Xiong et al., 2026). These methods provide strong evidence that future failures can be anticipated before an explicitly unsafe action appears.

RISA is most closely related to this predictive line, but couples forecasting and intervention differently. Instead of generating future continuations or producing only a current guard decision, RISA directly estimates the probability of a future violation at multiple horizons together with time-to-risk. The resulting signal represents both *risk magnitude* and *temporal urgency*. RISA then uses these predictions as the state of a learned intervention policy that determines when intervention is warranted and how strongly a safety-relevant representation direction should modify the acting model. This connects trajectory-level forecasting to adaptive representation-level control, allowing intervention strength to vary with the predicted evolution of risk rather than applying the same hard response to every flagged trajectory.