WHEN COLLABORATION BACKFIRES: PHANTOM-CONTROL FOR CREDIT PRESERVATION IN MULTI-AGENT LLMS

Anonymous authors
Paper under double-blind review

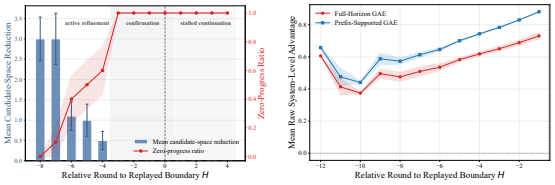
ABSTRACT

Multi-agent LLM systems use persistent shared memory to coordinate decisions across multiple rounds. However, continued collaboration can become harmful when locally unproductive updates persist across roles and rounds. We identify this failure mode as *Phantom Collaboration*, where a stalled collaborative tail can distort the credit assigned to earlier useful coordination. Our analysis shows that a sufficiently strong stalled-tail contribution can overwhelm the contribution of the pre-entry prefix and make the expected advantage of an earlier useful action negative. To address this problem, we propose PHANTOM-CONTROL, which detects persistent low observable progress and sets a common optimization boundary to restrict policy optimization to the retained collaborative prefix. At the ideal boundary, removing the stalled continuation preserves higher expected prefix credit. For the practical detector, we derive an exponential bound on the fixed-window false-trigger probability. Across five multi-agent active-reasoning tasks, PHANTOM-CONTROL achieves late-stage reward gains reaching 35–50% and reduces mean total generated response length by 20–35% in representative regimes. Additional controls indicate that the gains depend on targeted system-level boundary placement rather than trajectory shortening alone.

1 INTRODUCTION

Large language model (LLM) agents increasingly solve active-reasoning tasks through sequential interaction rather than one-shot generation Zhu et al. (2025); Park et al. (2025); Ke et al. (2026). In multi-agent systems, role-specific agents exchange intermediate results through persistent shared memory. This allows later roles and rounds to build on earlier interaction Zhang et al. (2025a; 2026); Guan et al. (2026). Such persistence is useful when continued interaction improves the shared collaborative state. But what happens to learning when system-level progress has stalled?

Once a locally unproductive update enters persistent shared memory, its influence need not remain local. It can affect later roles and future rounds. As a result, local low progress can persist throughout the collaborative trajectory. We call this failure mode *Phantom Collaboration*. The agents continue to interact, but their expected system-level progress is non-positive. The resulting continuation forms a stalled collaborative tail.



(a) Collaborative Stagnation (b) Temporal Credit Distortion

Figure 1: **Phantom Collaboration and its effect on temporal credit assignment.** (a) Collaborative progress diminishes around the detected boundary and remains stalled afterward. (b) Excluding the stalled continuation yields higher pre-boundary system-level advantage, consistent with temporal credit distortion. Results are shown on CD.

This stalled tail is not merely wasted interaction. Under outcome-reward RL Park et al. (2025); Feng et al. (2026); Venkataramani et al. (2026); Chen et al. (2026), later residuals can affect the advantage estimates of earlier actions through temporal credit assignment Kazemnejad et al. (2024);

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.

Feng et al. (2026); Nagpal et al. (2025); Chen et al. (2026). Shared memory carries stalled states forward across roles and rounds. Temporal credit assignment carries their training effects backward to earlier coordination. A sufficiently strong stalled-tail contribution can therefore suppress the expected advantage of an earlier useful action. It can even make the expected advantage negative. Figure 1 illustrates this mechanism. Collaborative progress stalls around the detected boundary. Excluding the stalled continuation preserves higher credit for earlier coordination.

This mechanism also explains why simply shortening interaction is not enough. A generic early-stopping rule may reduce trajectory length, but does not identify which continuation is responsible for distorting earlier credit Daley et al. (2023); Zou et al. (2026). Likewise, filtering an individual role output does not remove a stalled state that has already propagated through shared memory Wang et al. (2025b); Zhang et al. (2024), while reward shaping changes the learning signal itself rather than controlling which future residuals contribute to earlier decisions. The relevant object of intervention is therefore the collaborative trajectory over which temporal credit is assigned.

We therefore introduce PHANTOM-CONTROL. It is a system-level intervention that controls the trajectory used for temporal credit assignment. PHANTOM-CONTROL detects persistent low observable progress and sets a common optimization boundary. It retains the collaborative prefix through the confirmed boundary. Residuals from the subsequent stalled continuation are excluded from the credit assigned to earlier actions.

Our main contributions are:

- We identify and formalize *Phantom Collaboration*, a previously uncharacterized failure mode in shared-memory multi-agent systems. Locally unproductive updates can persist through shared memory across roles and rounds and form a stalled collaborative tail. This shows how local low progress can develop into system-level collaborative stagnation.
- To our knowledge, we provide the first formal analysis connecting persistent shared-memory collaborative stagnation to temporal credit distortion in multi-agent LLM reinforcement learning. Our analysis shows that a sufficiently strong stalled-tail contribution can overwhelm the pre-entry contribution. As a result, the expected advantage of an earlier useful action can become negative. Removing the ideal stalled continuation preserves higher expected prefix credit, while the practical detector admits an exponential bound on its fixed-window false-trigger probability.
- We propose PHANTOM-CONTROL to control which part of the collaborative trajectory contributes to temporal credit assignment. Across five interactive active-reasoning tasks, PHANTOM-CONTROL improves late-stage optimization. It achieves reward gains of 35–50% in representative regimes. It also reduces mean total generated response length by 20–35%. Random-Matched controls show that comparable trajectory shortening alone does not reproduce these gains, while Role-Local controls support shared-trajectory System-Level control over an estimator-local intervention.

2 RELATED WORK

Collaborative Reasoning Collaborative reasoning improves LLM agents through structured interaction rather than isolated generation. Zhang et al. (2024) distribute long-context inputs across sequential workers and use a manager to synthesize their messages. Qian et al. (2025) study how collaboration networks and topologies shape collective performance. Collaboration is also fragile: discussion can propagate incorrect answers Wang et al. (2024), while model interaction can expose knowledge gaps before unreliable generation Feng et al. (2024). These works primarily study how interaction structure affects collective reasoning behavior. We instead study a failure induced by persistence: shared memory can retain a locally unproductive update and propagate it across later roles and rounds, turning local low progress into system-level collaborative stagnation.

Credit Assignment Credit assignment remains a central obstacle in multi-turn LLM-agent training because sparse outcome rewards must be propagated through long interaction traces. Prior work uses turn-level intrinsic rewards to guide information acquisition Wan et al. (2026) or refines delayed feedback into process- and progress-level signals Zhang et al. (2025b); Yu et al. (2025); Wang et al. (2025a). Actor–critic training can further reduce variance under sparse rewards Liu et al. (2026). Belief deviation can also produce uninformative trajectory tails that contaminate credit assigned to

earlier actions Zou et al. (2026). Our setting introduces a distinct collaborative failure mechanism: a locally unproductive update can persist through shared memory across different roles and rounds, turning local stagnation into a system-level collaborative state whose continuation can jointly distort credit assigned to multiple role-specific decisions. PHANTOM-CONTROL addresses this failure through a common optimization boundary defined after complete system transitions while keeping policy updates role-specific.

3 MULTI-AGENT RL WITH SHARED MEMORY

We study sequential multi-agent active reasoning in which role-specific agents coordinate through shared memory. Each agent A_i has one functional role. We use *shared memory* to refer to collaborative context that persists across roles and rounds Huot et al. (2025). We consider n agents $A_1, \ldots, A_n$ interacting with an environment whose latent state space is S. The ground-truth state $s^\star \in S$ remains fixed within each episode. Let $\Delta(S)$ denote the probability simplex over S. At the beginning of round t, the system state is

$$X_t = (B_t, b_t^{\mathrm{mem}}, M_t), \qquad B_t = (b_{1,t}, \ldots, b_{n,t}),$$

where $b_{i,t}$ is the belief of agent A_i, b_t^{mem} is the belief induced by shared memory, and $M_t \in \mathbb{M}$ is the current shared-memory content. Here $\mathbb{M}$ denotes the shared-memory content space, and the fixed and non-trainable shared-memory mechanism is denoted by $\mathcal{M}$. The belief components are analytical variables and are not directly observed by the acting roles. For the analysis, we define the belief projection as $\bar{X}_t := \mathrm{proj}_{\mathrm{bel}}(X_t) = (B_t, b_t^{\mathrm{mem}})$. When a policy is written as conditioned on X_t, this means that the role conditions on the actor-observable shared-memory history represented within the system state.

Within each round, the roles generate proposals sequentially in the fixed order $1, \ldots, n$. Let $\pi_{\theta_i}^{\mathrm{tok}}$ denote the autoregressive token policy of role i, and let $\pi_{\theta_i}^{\mathrm{prop}}$ denote the induced distribution over its semantic proposal. Let $\Theta = (\theta_1, \ldots, \theta_n)$. Role i generates

$$u_{i,t} \sim \pi_{\theta_i}^{\mathrm{prop}}(\cdot \mid X_t, u_{<i,t}), \qquad u_{<i,t} := (u_{1,t}, \ldots, u_{i-1,t}),$$

where $u_{<1,t}$ is empty. After all roles have generated their proposals, we write $U_t = (u_{1,t}, \ldots, u_{n,t})$. The variable $u_{i,t}$ denotes the semantic proposal of role i. Its token realization is denoted by $Y_{i,t}$ when we define the token-wise policy objective in Appendix F. The ordered proposals are then mapped to one environment-facing action. The fixed shared-memory action-selection kernel gives $a_t^{\mathrm{env}} \sim \kappa_{\mathcal{M}}(\cdot \mid X_t, U_t)$. The kernel $\kappa_{\mathcal{M}}$ is non-trainable and is not an additional centralized actor. Let $\mathcal{O}(\cdot \mid s, a)$ denote the environment observation kernel. The environment then returns $o_t \sim \mathcal{O}(\cdot \mid s^\star, a_t^{\mathrm{env}})$.

Across rounds, information stored in the updated system state can affect later decisions through shared memory. A locally unproductive update therefore need not remain local. It can affect later roles and persist into later rounds. We next characterize when this persistence develops into system-level collaborative stagnation.

4 PHANTOM COLLABORATION: MECHANISM AND CREDIT CONSEQUENCES

We first characterize shared-memory-mediated collaborative stagnation and then analyze its effect on system-level credit assignment. The collaborative stagnation region (CSR) is a latent high-potential region with non-positive expected system-level progress. *Phantom Collaboration* denotes the persistent regime after CSR entry. We first introduce local agent stagnation and then show how it can persist through shared memory and develop into system-level collaborative stagnation.

4.1 SHARED-MEMORY-MEDIATED COLLABORATIVE STAGNATION

Local agent stagnation. We first introduce local agent stagnation. To describe progress at the agent level, we use the ground-truth log-loss potential $\Psi_{i,t} := -\log b_{i,t}(s^\star)$, where $b_{i,t}$ is the belief of agent A_i at round t. A smaller value of $\Psi_{i,t}$ indicates greater confidence in the true state. We say that agent A_i enters local stagnation when $\Psi_{i,t+1} \geq \Psi_{i,t}$. This means that the agent does not make

progress over that round. Local stagnation alone is not a system-level failure. It becomes relevant at the system level when the low-progress state is preserved through shared memory and influences later roles and later rounds.

Shared-memory-mediated propagation. Once a low-progress update enters shared memory, its effect can extend beyond the role that generated it. It may affect later roles in the same round and continue to influence later rounds. Agent-level progress alone is therefore not enough to describe the state of the collaborative system.

We measure system-level progress by combining memory-level and agent-level potentials. The memory-level potential is $\Psi_t^{\mathrm{mem}} := -\log b_t^{\mathrm{mem}}(s^\star)$. Together with the agent-level potentials, it defines the system-level potential

$$\Psi^{\mathrm{sys}}(\bar{X}_t) := \alpha\Psi_t^{\mathrm{mem}} + (1-\alpha)\sum_{i=1}^{n} w_i\Psi_{i,t}, \quad \alpha \in [0,1], \quad w_i \geq 0, \quad \sum_{i=1}^{n} w_i = 1,$$

where $\bar{X}_t = (B_t, b_t^{\mathrm{mem}})$ is the belief projection of the system state. For brevity, we also write $\Psi^{\mathrm{sys}}(X_t) := \Psi^{\mathrm{sys}}(\bar{X}_t)$. The parameter α controls the weight of the memory-level component, while w_i controls the contribution of each role. Because the system-level potential includes the belief induced by shared memory, progress by one role does not necessarily imply progress of the collaborative system.

The system-level potential allows us to describe whether the collaborative system improves from one round to the next. We condition on the interaction history available before the round-t action is generated. This history defines the pre-action filtration $\mathcal{F}_t^{\mathrm{act}}$. The one-step expected progress of the actual shared-memory system is then

$$\mathcal{P}^{\mathrm{sys}}_{\Theta,\mathcal{M}}(X_t) := \Psi^{\mathrm{sys}}(X_t) - \mathbb{E}\left[\Psi^{\mathrm{sys}}(X_{t+1}) \mid \mathcal{F}_t^{\mathrm{act}}\right].$$

When $\mathcal{P}^{\mathrm{sys}}_{\Theta,\mathcal{M}}(X_t) \leq 0$, the expected system-level potential does not decrease. The system may therefore continue to collaborate without making positive expected system-level progress.

This condition describes when progress stops, but it does not explain how the collaborative update contributes to this behavior. To make this comparison, we use a Bayesian oracle as a theoretical benchmark. The oracle uses the same environment-facing action and observation as the actual system, but updates each belief component by Bayes' rule. For a belief b, environment-facing action a^{env}, and observation o, the Bayesian update is

$$\mathcal{B}^\star_{\mathrm{Bayes}}(b, a^{\mathrm{env}}, o)(s) := \frac{\mathcal{O}(o \mid s, a^{\mathrm{env}})b(s)}{\sum_{s' \in S}\mathcal{O}(o \mid s', a^{\mathrm{env}})b(s')}, \qquad s \in S.$$

For $\bar{X}_t = (B_t, b_t^{\mathrm{mem}})$ with $B_t = (b_{1,t}, \ldots, b_{n,t})$, we apply this Bayesian update to every agent belief and to the memory-induced belief. This gives $\bar{\mathfrak{B}}^\star(\bar{X}_t, a_t^{\mathrm{env}}, o_t) := (B_t^+, b_t^{\mathrm{mem},+})$, where $b_{i,t}^+ = \mathcal{B}^\star_{\mathrm{Bayes}}(b_{i,t}, a_t^{\mathrm{env}}, o_t)$ and $b_t^{\mathrm{mem},+} = \mathcal{B}^\star_{\mathrm{Bayes}}(b_t^{\mathrm{mem}}, a_t^{\mathrm{env}}, o_t)$. This belief-only oracle update is used only for the analysis and is not available to the acting roles.

We can now compare the actual collaborative update with the Bayesian update under the same action and observation. The one-step actual–oracle potential gap is

$$c^{\mathrm{sys}}_{\Theta,\mathcal{M}}(X_t) := \mathbb{E}\left[\Psi^{\mathrm{sys}}\left(\mathfrak{B}_\Theta(X_t, U_t, a_t^{\mathrm{env}}, o_t)\right) - \Psi^{\mathrm{sys}}\left(\bar{\mathfrak{B}}^\star(\bar{X}_t, a_t^{\mathrm{env}}, o_t)\right) \,\middle|\, \mathcal{F}_t^{\mathrm{act}}\right].$$

A positive value means that the actual update leaves the system at a higher potential than the corresponding Bayesian update. We next connect this gap to system-level stagnation. Proposition 4.1 shows that, when the actual–oracle gap grows with the system-level potential, sufficiently high-potential states have non-positive expected system-level progress.

Proposition 4.1 (Informal collaborative stagnation region). *Suppose that, for sufficiently high-potential states, $c^{\mathrm{sys}}_{\Theta,\mathcal{M}}(X) \geq m_{\Theta,\mathcal{M}}\Psi^{\mathrm{sys}}(X) - c_0^{\mathrm{sys}}$, where $m_{\Theta,\mathcal{M}} > 0$ and $c_0^{\mathrm{sys}} \geq 0$. Then there exists a threshold Ψ_{CSR} such that $\Psi^{\mathrm{sys}}(X) \geq \Psi_{\mathrm{CSR}}$ implies $\mathcal{P}^{\mathrm{sys}}_{\Theta,\mathcal{M}}(X) \leq 0$ for every reachable state X. If the region $\mathcal{R}^{\mathrm{sys}}_{\Theta,\mathcal{M}}(\Psi_{\mathrm{CSR}}) := \{X : \Psi^{\mathrm{sys}}(X) \geq \Psi_{\mathrm{CSR}}\}$ is absorbing under the actual policy-averaged dynamics, then it is a collaborative stagnation region (CSR). Hence, if $t_S^{\mathrm{sys}} := \inf\{t \geq 1 : \Psi_t^{\mathrm{sys}} \geq \Psi_{\mathrm{CSR}}\} < \infty$, then for all $t \geq t_S^{\mathrm{sys}}$, $\mathcal{P}^{\mathrm{sys}}_{\Theta,\mathcal{M}}(X_t) \leq 0$, equivalently $\mathbb{E}[\Psi_{t+1}^{\mathrm{sys}} \mid \mathcal{F}_t^{\mathrm{act}}] \geq \Psi_t^{\mathrm{sys}}$.*

Proposition 4.1 identifies a regime in which the system continues to collaborate but no longer makes positive expected system-level progress. Continued interaction can therefore extend a stalled tail instead of correcting the collaborative state. We next show that this stalled continuation is more than an efficiency problem. In outcome-reward RL, it can also distort credit assigned before CSR entry.

4.2 System-Level Credit Distortion and Advantage Inversion

Once the system enters a CSR, the stalled continuation can affect the credit assigned to actions taken before entry. The reason is that full-horizon generalized advantage estimation (GAE) propagates later TD residuals backward to earlier actions. Let t_S^{sys} denote the first CSR hitting time and let T be the terminal state index. We condition on $\mathsf{E}_\tau := \{t_S^{\mathrm{sys}} = \tau\}$, where $\tau \in \{2, \dots, T-1\}$. For any $t < \tau$, the action at round t belongs to the pre-entry prefix, but its full-horizon advantage still contains TD residuals generated after CSR entry. This is the source of system-level credit distortion.

To make this backward effect explicit, we write the full-horizon GAE in terms of system-level TD residuals. The centralized system-level critic is $V_t^{\mathrm{sys}} := V^{\mathrm{sys}}(X_t)$, and r_t denotes the realized environment reward at round t. The corresponding TD residual is

$$\delta_t^{\mathrm{sys}} := r_t + \gamma V_{t+1}^{\mathrm{sys}} - V_t^{\mathrm{sys}}, \qquad t = 1, \dots, T-1,$$

where γ is the reward discount factor. The full-horizon system-level GAE, $\widehat{A}_t^{\mathrm{sys}}$, is used as the common credit signal for the role-specific policy updates. With GAE trace parameter λ, its value for $t < \tau$ can be written as

$$\widehat{A}_t^{\mathrm{sys}} = \underbrace{\sum_{j=0}^{\tau-t-1} (\gamma\lambda)^j \delta_{t+j}^{\mathrm{sys}}}_{\text{pre-entry contribution}} + \underbrace{\sum_{j=\tau-t}^{T-t-2} (\gamma\lambda)^j \delta_{t+j}^{\mathrm{sys}}}_{\text{nonterminal stalled tail}} + \underbrace{(\gamma\lambda)^{T-t-1} \delta_{T-1}^{\mathrm{sys}}}_{\text{terminal residual}}.$$

This decomposition makes the source of the distortion explicit. The first term contains credit from the pre-entry prefix, while the second contains TD residuals generated during the post-entry stalled continuation. The terminal residual is kept separate because it is not part of the nonterminal stalled tail. We use value and credit regularity conditions to connect post-entry degradation of the collaborative state to negative post-entry value drift. These conditions allow the stalled-tail contribution to be compared directly with the pre-entry contribution.

Define the ideal pre-entry estimator and its conditional expected credit by

$$\widehat{A}_{t,\mathrm{pre}}^{\mathrm{sys}} := \sum_{j=0}^{\tau-t-1} (\gamma\lambda)^j \delta_{t+j}^{\mathrm{sys}}, \quad \mu_{\mathrm{pre}}^{\mathrm{sys}}(t,\tau) := \mathbb{E}\left[\widehat{A}_{t,\mathrm{pre}}^{\mathrm{sys}} \mid \mathsf{E}_\tau\right], \quad \Gamma_{\mathrm{tail}}^{\mathrm{sys}}(t,\tau) := \sum_{j=\tau-t}^{T-t-2} (\gamma\lambda)^j,$$

where an empty sum is zero. The quantity $\mu_{\mathrm{pre}}^{\mathrm{sys}}(t,\tau)$ is the ideal pre-entry estimator's conditional expected credit, and $\Gamma_{\mathrm{tail}}^{\mathrm{sys}}(t,\tau)$ is the discounted mass of the nonterminal stalled tail.

Theorem 4.2 (**Informal system-level credit degradation and advantage inversion**). *Condition on $\mathsf{E}_\tau = \{t_S^{\mathrm{sys}} = \tau\}$ with $\mathbb{P}(\mathsf{E}_\tau) > 0$, and fix $t < \tau$. Under Proposition 4.1 and the stated value and credit regularity conditions, suppose that the post-entry system truth-confidence $\chi_t^{\mathrm{sys}} := \exp(-\Psi_t^{\mathrm{sys}})$ has negative drift of magnitude $\rho_\chi^{\mathrm{sys}} > 0$, and that the system-level critic satisfies $V_t^{\mathrm{sys}} = g(\chi_t^{\mathrm{sys}})$ with $g'(x) \geq \kappa_V > 0$. Then $\mathbb{E}[\widehat{A}_{t,\mathrm{pre}}^{\mathrm{sys}} - \widehat{A}_t^{\mathrm{sys}} \mid \mathsf{E}_\tau] \geq \gamma\kappa_V \rho_\chi^{\mathrm{sys}} \Gamma_{\mathrm{tail}}^{\mathrm{sys}}(t,\tau)$. Thus, the stalled continuation lowers the expected credit assigned to pre-entry coordination by at least $\gamma\kappa_V \rho_\chi^{\mathrm{sys}} \Gamma_{\mathrm{tail}}^{\mathrm{sys}}(t,\tau)$. In particular, if $0 < \mu_{\mathrm{pre}}^{\mathrm{sys}}(t,\tau) < \gamma\kappa_V \rho_\chi^{\mathrm{sys}} \Gamma_{\mathrm{tail}}^{\mathrm{sys}}(t,\tau)$, then $\mathbb{E}[\widehat{A}_t^{\mathrm{sys}} \mid \mathsf{E}_\tau] < 0$.*

Prior multi-agent studies have mainly characterized neighboring aspects of this problem empirically: erroneous agent outputs can propagate through communication structures and affect collective decisions Shen et al. (2025), while team outcomes can be decomposed into agent-specific credit Nagpal et al. (2025). Our theorem formalizes how these phenomena can become an optimization failure in shared-memory collaboration. It proves that a post-entry stalled continuation lowers the common system-level credit assigned to earlier coordination by at least $\gamma\kappa_V \rho_\chi^{\mathrm{sys}} \Gamma_{\mathrm{tail}}^{\mathrm{sys}}(t,\tau)$, and gives a sufficient condition under which a positive ideal-prefix advantage becomes negative. Because this common signal drives the role-specific policy updates, the distortion can affect multiple roles simultaneously. The result therefore establishes a quantitative link between multi-agent information propagation and temporal credit distortion, motivating the common optimization boundary introduced next.

5 PHANTOM-CONTROL: CREDIT-PRESERVING CONTROL OF STALLED COLLABORATIVE TAILS

The previous section shows that a stalled continuation can distort the credit assigned to earlier coordination. PHANTOM-CONTROL is designed to prevent this distortion. We first describe how persistent low observable progress is detected at the system level.

5.1 OBSERVABLE PROGRESS DETECTION

The system-level potential Ψ_t^{sys} and the latent CSR entry time t_S^{sys} are used in the theoretical analysis but are not directly observed during rollout generation. The detector instead uses an observable progress proxy computed after each complete system transition. The realized system-level progress is $G_t^{\mathrm{sys}} := \Psi_t^{\mathrm{sys}} - \Psi_{t+1}^{\mathrm{sys}}$. It describes the change in the system-level potential over one transition, but remains latent because Ψ_t^{sys} is not directly observed. Let $\mathcal{H}_t^{\mathrm{sys}}$ denote the observable system trace available at the start of round t. After round t, the detector computes $D_t^{\mathrm{sys}} := D_{\mathrm{obs}}\big(\mathcal{H}_t^{\mathrm{sys}}, \mathcal{H}_{t+1}^{\mathrm{sys}}\big)$. The score D_t^{sys} is the observable progress proxy for G_t^{sys}. The three quantities describe progress at different levels. The quantity $\mathcal{P}_{\Theta,\mathcal{M}}^{\mathrm{sys}}(X_t)$ is conditional expected progress, G_t^{sys} is latent realized progress, and D_t^{sys} is observable proxy progress.

However, a single low value of D_t^{sys} may result from noise. The detector therefore requires low progress to persist over k consecutive system transitions. For threshold $\Delta_{\min}$ and window length k, define

$$\mathcal{W}_{t,k}^{\mathrm{sys}} := \bigcap_{j=0}^{k-1} \left\{ D_{t+j}^{\mathrm{sys}} \le \Delta_{\min} \right\}, \qquad 1 \le t \le T-k,$$

where T is the terminal state index. The event $\mathcal{W}_{t,k}^{\mathrm{sys}}$ means that low observable progress persists throughout the window. The following result explains why the detector requires persistent low progress rather than a single low-progress transition.

Proposition 5.1 (Informal fixed-window false-trigger control). *Consider a prescribed informative window of length k. Suppose that, for $j = 0, \dots, k-1$, $G_{t+j}^{\mathrm{sys}} \ge \rho > 0$ and $D_{t+j}^{\mathrm{sys}} = G_{t+j}^{\mathrm{sys}} + \beta_{t+j}^{\mathrm{sys}} + \xi_{t+j}^{\mathrm{sys}}$, with $|\beta_{t+j}^{\mathrm{sys}}| \le M_d$. Suppose also that $\xi_t^{\mathrm{sys}}, \dots, \xi_{t+k-1}^{\mathrm{sys}}$ are independent, centered, σ-sub-Gaussian noise variables for some $\sigma > 0$. If $\Delta_{\min} < \rho - M_d$, then $\mathbb{P}(\mathcal{W}_{t,k}^{\mathrm{sys}}) \le \exp\big(-k(\rho - M_d - \Delta_{\min})^2/(2\sigma^2)\big)$.*

Proposition 5.1 shows that the fixed-window false-trigger probability decreases exponentially with k when informative progress remains separated from the threshold. This explains why the detector requires persistent low progress rather than a single low-progress transition. The proposition applies to a fixed window. The first confirmed low-progress window is then used to set the common optimization boundary described next.

5.2 COMMON OPTIMIZATION BOUNDARY

The common optimization boundary is

$$\widehat{H} := \inf\left(\left\{ t+k : 1 \le t \le T-k,\ \mathcal{W}_{t,k}^{\mathrm{sys}} \text{ holds} \right\} \cup \{T\} \right).$$

If no window triggers, $\widehat{H} = T$. Otherwise, the first confirmed low-progress window sets the boundary at its end. The method retains the confirmation window and $X_{\widehat{H}}$. It excludes only the post-confirmation continuation. The practical boundary $\widehat{H}$ need not coincide with the latent CSR entry time t_S^{sys}.

5.3 PREFIX-SUPPORTED CREDIT ASSIGNMENT

Let H denote the common optimization boundary, with $H = \widehat{H}$ in practice. Policy-gradient updates use only the retained collaborative prefix ending at X_H. All roles use the same prefix mask,

$$m_t^{\mathrm{PC}} := \mathbf{1}\{t < H\} = \mathbf{1}\{t < \widehat{H}\}, \quad t = 1, \dots, T-1.$$

The mask selects complete system transitions and retains all role-specific decisions within them. The boundary H determines which future TD residuals enter the advantage estimates of those decisions. The reward itself is unchanged, and no additional process reward is introduced. Because the mask uses the strict inequality $t < H$, the final retained transition is the one ending at X_H. Decisions with $t \geq H$ receive zero policy-gradient weight.

Using the system-level TD residuals δ_t^{sys} defined in Section 4.2, PHANTOM-CONTROL replaces full-horizon GAE with the prefix-supported estimator,

$$\widehat{A}_{t,<H}^{\mathrm{sys}} := \begin{cases} \sum_{\ell=t}^{H-1} (\gamma\lambda)^{\ell-t} \delta_\ell^{\mathrm{sys}}, & t < H, \\ 0, & t \geq H, \end{cases} \qquad t = 1, \ldots, T-1.$$

Restricting the GAE horizon from T to H prevents residuals after the boundary from entering the advantage estimates of retained actions. For a boundary $H < T$, the last retained residual remains

$$\delta_{H-1}^{\mathrm{sys}} = r_{H-1} + \gamma V^{\mathrm{sys}}(X_H) - V^{\mathrm{sys}}(X_{H-1}).$$

An artificial boundary $H < T$ is an optimization boundary and not an environment terminal state. Therefore, $V^{\mathrm{sys}}(X_H)$ is retained for bootstrapping. It is set to zero only when X_H is terminal. For the theoretical analysis, the ideal boundary is the CSR hitting time t_S^{sys}. Conditional on E_τ, this gives $H = \tau$.

Corollary 5.2 (Informal ideal-boundary credit preservation). *Under the regularity assumptions of Theorem 4.2, condition on $\mathsf{E}_\tau = \{t_S^{\mathrm{sys}} = \tau\}$ with $\mathbb{P}(\mathsf{E}_\tau) > 0$. Set the ideal boundary $H = \tau$ and fix $t < \tau$. Recall $\widehat{A}_{t,\mathrm{pre}}^{\mathrm{sys}} := \sum_{\ell=t}^{\tau-1} (\gamma\lambda)^{\ell-t} \delta_\ell^{\mathrm{sys}}$. Then*

$$\mathbb{E}\left[\widehat{A}_{t,\mathrm{pre}}^{\mathrm{sys}} \,\middle|\, \mathsf{E}_\tau\right] \geq \mathbb{E}\left[\widehat{A}_t^{\mathrm{sys}} \,\middle|\, \mathsf{E}_\tau\right] + \gamma\kappa_V \rho_\chi^{\mathrm{sys}} \Gamma_{\mathrm{tail}}^{\mathrm{sys}}(t, \tau).$$

Corollary 5.2 shows that the ideal boundary removes the negative nonterminal tail contribution. The expected credit assigned to the retained prefix is therefore higher. For the practical detector, Proposition 5.1 bounds the false-trigger probability for a prescribed informative window.

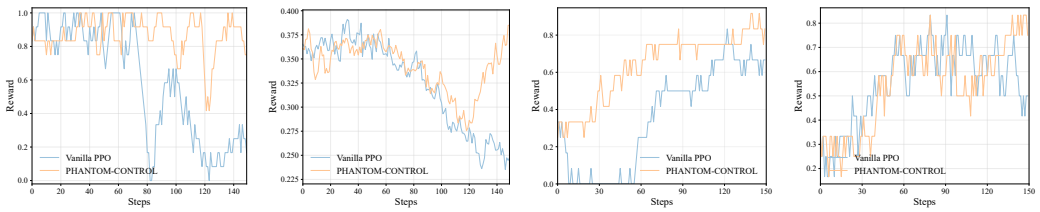
6 EXPERIMENTS

We evaluate PHANTOM-CONTROL on five interactive active-reasoning tasks: *Circuit Decoding (CD), Situation Puzzles (SP), Guess Numbers (GN), Preference Estimation (PE)*, and *Movie Recommendation (MR)*. All methods are evaluated within the same three-role shared-memory system and trained with proximal policy optimization (PPO). We report reward and generated-response-length results to characterize the effectiveness and efficiency of PHANTOM-CONTROL. We also use control experiments to test whether its gains can be explained by trajectory shortening alone or instead depend on where the optimization boundary is placed. Complete implementation details and additional experiments are provided in Sections E–H of the Appendix.

6.1 DOES CONTROLLING STALLED COLLABORATION IMPROVE OPTIMIZATION?

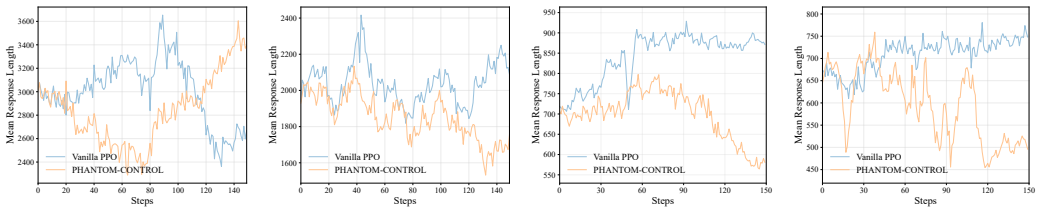
Figures 2 and 3 summarize the reward and generated-response-length dynamics, respectively. The reward advantage of PHANTOM-CONTROL emerges primarily during the middle and late stages of training. On CD, it maintains an absolute advantage of about 0.6–0.8 after Vanilla PPO begins to degrade. SP and PE show sustained late-stage separation, with an approximately 35–50% advantage on SP during the final 40–50 steps. On MR, the separation begins earlier and persists thereafter. The timing of these gains reveals a shared pattern across tasks: PHANTOM-CONTROL becomes most useful when optimization begins to slow down or lose previously acquired performance. This is consistent with its intended mechanism. Once progress stalls, limiting the influence of post-boundary residuals can prevent later updates from eroding useful earlier behavior and can stabilize subsequent optimization.

The response-length results reveal a related behavioral effect. On SP, MR, and PE, the higher late-stage rewards under PHANTOM-CONTROL coincide with reductions of roughly 20–35% in mean total generated response length. Thus, targeted intervention can improve both optimization stability



(a) CD (b) SP (c) MR (d) PE

Figure 2: **Training reward dynamics under Vanilla PPO and PHANTOM-CONTROL.** The x-axis denotes PPO outer training steps, and the y-axis reports the corresponding training reward.



(a) CD (b) SP (c) MR (d) PE

Figure 3: **Mean total generated response length under Vanilla PPO and PHANTOM-CONTROL.** The x-axis denotes PPO outer training steps, and the y-axis reports, for each rollout, the total number of tokens generated by all roles across all rounds, averaged over the evaluated rollouts.

and the efficiency of the resulting behavior. However, the relationship between reward and response length is not uniform across tasks. Response shortening is therefore an incomplete explanation for the reward gains. The more central question concerns how the boundary changes the trajectory information used for policy updates. The following controls address this question by examining when the boundary activates, whether the amount or placement of tail masking drives the gains, and whether shared-trajectory System-Level control is more effective than an estimator-local intervention.

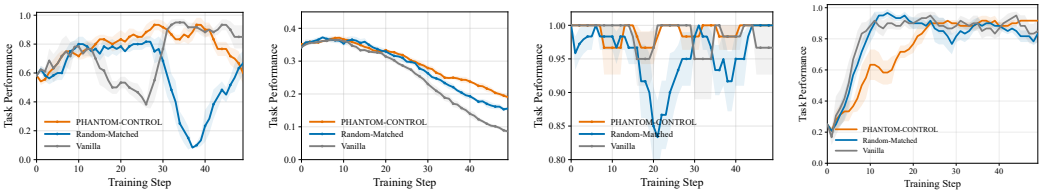
6.2 How Does Persistence Affect the Intervention?

For the persistence analysis, let Q_t denote the task-specific evaluation score at state X_t. Boundary Activation Rate is the fraction of evaluated rollouts with $\widehat{H} < T$. On the k-specific activated subset, Masked Tail Fraction is $(T - \widehat{H})/(T - 1)$, Boundary Score is $Q_{\widehat{H}}$, and Positive-Gain Rate is the fraction with $Q_T > Q_{\widehat{H}}$. Except for Boundary Activation Rate, these statistics are conditional on activation and therefore use different subsets across values of k.

Table 1: **Effect of persistence on boundary activation and post-boundary progress.** GN and PE are shown as illustrative cases. Boundary Score is reported as mean $\pm$ SEM over scored activated trajectories pooled across training checkpoints.

Task	k	Boundary Activation Rate	Masked Tail Fraction	Boundary Score	Positive-Gain Rate
GN	1	87.7%	44.4%	0.496 ± 0.008	49.0%
	2	59.8%	17.3%	0.814 ± 0.008	16.7%
	3	32.6%	6.3%	0.922 ± 0.007	6.4%
	4	13.3%	1.9%	0.959 ± 0.009	3.4%
PE	3	97.1%	56.6%	0.302 ± 0.007	8.9%
	6	65.9%	19.7%	0.350 ± 0.009	5.5%
	8	47.3%	4.7%	0.348 ± 0.011	3.0%

Table 1 shows that increasing k makes boundary placement more conservative: the detector activates less often and, when activated, masks a smaller part of the trajectory. On GN, for example, increasing k from 1 to 4 reduces the Boundary Activation Rate from 87.7% to 13.3%; the Positive-Gain Rate also decreases with k, with PE exhibiting the same overall trend. These patterns are consistent with the intended role of persistence. A small k can react to a brief period of low progress and place the boundary while useful improvement may still follow, whereas a larger k requires low progress to persist over more consecutive transitions, resulting in less frequent, later intervention and smaller masked tails. Thus, persistence helps distinguish a temporary slowdown from a sustained stalled continuation rather than simply maximizing the amount of masked trajectory. Because each value of k induces a different activated subset, these comparisons should



(a) CD (b) SP (c) GN (d) PE

Figure 4: **PHANTOM-CONTROL versus Random-Matched.** The x-axis denotes PPO outer training steps, and the y-axis reports the corresponding task-specific evaluation metric. The solid curves show the overall training trend, while the shaded regions indicate local variation over a three-step window.

be interpreted as sensitivity results rather than paired causal effects. Appendix G further shows how boundary activation evolves across different persistence settings during training.

6.3 DOES OPTIMIZATION-BOUNDARY PLACEMENT MATTER?

At each PPO outer step, Random-Matched uses the same number of activated rollouts as PHANTOM-CONTROL. It also uses the same round-level masked-tail-length multiset. However, it assigns the matched boundaries randomly across rollouts. The control therefore preserves the intervention frequency and the distribution of masking amounts. What it removes is the correspondence between each boundary and the stalled continuation detected in that rollout.

On CD and GN, Random-Matched exhibits pronounced middle-to-late-stage degradation, whereas PHANTOM-CONTROL remains substantially more stable. SP also shows a persistent late-stage advantage for PHANTOM-CONTROL. These differences cannot be attributed to intervention frequency or masking amount. At every PPO outer step, Random-Matched preserves both the activated rollout count and the complete masked-tail-length multiset across the rollout batch. If generic tail masking were sufficient, Random-Matched should produce similar training dynamics. Instead, randomly reassigning the same masking budget substantially weakens the result. This indicates that the benefit comes from conditioning the boundary on the progress observed within each rollout. Under PHANTOM-CONTROL, the boundary is placed where the detector identifies a stalled continuation. Random reassignment breaks this connection and can apply masking to a still-useful continuation while retaining a stalled continuation elsewhere. PE provides a less discriminative case because the two methods remain closer after the early stage, but the controlled separations on CD, GN, and SP consistently support the same mechanism. Overall, the comparison shows that the gains depend on where tail masking is applied, rather than on the amount of masking alone.

Random-Matched establishes that where the boundary is placed within a rollout matters. A complementary Role-Local control in Appendix H contrasts the shared-trajectory System-Level intervention with an estimator-local intervention. The comparison tests whether acting on the complete shared trajectory yields more persistent gains than intervening only at the selected local update.

7 CONCLUSION

This work shows that persistent collaboration can affect learning beyond the value of the additional interaction itself. When collaborative progress stalls, the remaining trajectory may still contribute to temporal credit assignment. These later contributions can weaken the learning signal of earlier useful coordination and, under sufficiently adverse conditions, make its expected advantage negative. Based on this observation, we develop PHANTOM-CONTROL to separate stalled continuation from useful collaborative credit. PHANTOM-CONTROL places a common optimization boundary on the shared trajectory and removes post-boundary residuals from the credit of retained actions. These findings suggest that persistent multi-agent learning should consider not only whether further interaction is useful, but also how it affects the learning signal of earlier decisions. In future work, we will further investigate finer-grained system-level credit assignment beyond hard boundary-based control, where different parts of a collaborative trajectory can contribute differently to earlier decisions.

AI USE STATEMENT

We developed the core research ideas and implemented the proposed method. We used generative AI tools for language editing and manuscript organization. We also used these tools to review mathematical claims and arguments and to check assumptions, notation, consistency, and implementation details. We reviewed and revised AI-assisted suggestions before incorporating them into the manuscript and take responsibility for the final text, mathematical claims, implementation, and experimental results.

REFERENCES

Kartikeya Badola, Jonathan Simon, Arian Hosseini, Sara Marie Mc Carthy, Tsendsuren Munkhdalai, Abhimanyu Goyal, Tomáš Kočiskỳ, Shyam Upadhyay, Bahare Fatemi, and Mehran Kazemi. Multi-turn puzzles: Evaluating interactive reasoning and strategic dialogue in llms. *arXiv preprint arXiv:2508.10142*, 2025.

Mengzhuo Chen, Junjie Wang, Fangwen Mu, Yawen Wang, Zhe Liu, Huanxiang Feng, and Qing Wang. Seeing the whole elephant: A benchmark for failure attribution in llm-based multi-agent systems. *arXiv preprint arXiv:2604.22708*, 2026.

Brett Daley, Martha White, Christopher Amato, and Marlos C Machado. Trajectory-aware eligibility traces for off-policy reinforcement learning. In *International Conference on Machine Learning*, pp. 6818–6835. PMLR, 2023.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for llm agent training. *Advances in Neural Information Processing Systems*, 38:46375–46408, 2026.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14664–14690, 2024.

Yao Guan, Lin Wang, Zhihu Lu, Ziyi Wang, Wenzhu Yan, and Qiang Duan. Moc: Multi-order communication in llm-based multi-agent systems. *arXiv preprint arXiv:2606.02359*, 2026.

Fantine Huot, Reinald Kim Amplayo, Jennimaria Palomaki, Alice Shoshana Jakobovits, Elizabeth Clark, and Mirella Lapata. Agents' room: Narrative generation through multi-step collaboration. In *International Conference on Learning Representations*, volume 2025, pp. 5150–5183, 2025.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. Vineppo: Refining credit assignment in rl training of llms. *arXiv preprint arXiv:2410.01679*, 2024.

Zixuan Ke, Yifei Ming, Austin Xu, Ryan Chin, Xuan-Phi Nguyen, Prathyusha Jwalapuram, Jiayu Wang, Semih Yavuz, Caiming Xiong, and Shafiq Joty. Mas-orchestra: Understanding and improving multi-agent reasoning through holistic orchestration and controlled benchmarks. *arXiv preprint arXiv:2601.14652*, 2026.

Shuo Liu, Tianle Chen, Ryan Amiri, and Christopher Amato. Learning decentralized llm collaboration with multi-agent actor critic. *arXiv preprint arXiv:2601.21972*, 2026.

Kartik Nagpal, Dayi Dong, Jean-Baptiste Bouvier, and Negar Mehr. Leveraging large language models for effective and explainable multi-agent credit assignment. *arXiv preprint arXiv:2502.16863*, 2025.

Chanwoo Park, Seungju Han, Xingzhi Guo, Asuman E Ozdaglar, Kaiqing Zhang, and Joo-Kyung Kim. Maporl: Multi-agent post-co-training for collaborative large language models with reinforcement learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30215–30248, 2025.

Chen Qian, Zihao Xie, Yifei Wang, Wei Liu, Kunlun Zhu, Hanchen Xia, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, et al. Scaling large language model-based multi-agent collaboration. In *International Conference on Learning Representations*, volume 2025, pp. 41488–41505, 2025.

Xu Shen, Yixin Liu, Yiwei Dai, Yili Wang, Rui Miao, Yue Tan, Shirui Pan, and Xin Wang. Understanding the information propagation effects of communication topologies in llm-based multi-agent systems. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12358–12372, 2025.

Vishal Venkataramani, Haizhou Shi, Zixuan Ke, Austin Xu, Xiaoxiao He, Yingbo Zhou, Semih Yavuz, Hao Wang, and Shafiq Joty. Mas-prove: Understanding the process verification of multi-agent systems. *arXiv preprint arXiv:2602.03053*, 2026.

Yanming Wan, Jiaxing Wu, Marwa Abdulhai, Lior Shani, and Natasha Jaques. Enhancing personalized multi-turn dialogue with curiosity reward. *Advances in Neural Information Processing Systems*, 38:155857–155894, 2026.

Hanlin Wang, Chak Tou Leong, Jiashuo Wang, Jian Wang, and Wenjie Li. Spa-rl: Reinforcing llm agents via stepwise progress attribution. *arXiv preprint arXiv:2505.20732*, 2025a.

Haochun Wang, Sendong Zhao, Jingbo Wang, Zewen Qiang, Bing Qin, and Ting Liu. Beyond frameworks: Unpacking collaboration strategies in multi-agent systems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21361–21375, 2025b.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. Rethinking the bounds of llm reasoning: Are multi-agent discussions the key? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6106–6131, 2024.

Haofei Yu, Zhengyang Qi, Yining Zhao, Kolby Nottingham, Keyang Xuan, Bodhisattwa Prasad Majumder, Hao Zhu, Paul Pu Liang, and Jiaxuan You. Sotopia-rl: Reward design for social intelligence. *arXiv preprint arXiv:2508.03905*, 2025.

Guibin Zhang, Yanwei Yue, Zhixun Li, Sukwon Yun, Guancheng Wan, Kun Wang, Dawei Cheng, Jeffrey Yu, and Tianlong Chen. Cut the crap: An economical communication pipeline for llm-based multi-agent systems. In *International Conference on Learning Representations*, volume 2025, pp. 75389–75428, 2025a.

Guibin Zhang, Muxin Fu, Kun Wang, Frank Wan, Miao Yu, and Shuicheng Yan. G-memory: Tracing hierarchical memory for multi-agent systems. *Advances in Neural Information Processing Systems*, 38:12988–13018, 2026.

Yusen Zhang, Ruoxi Sun, Yanfei Chen, Tomas Pfister, Rui Zhang, and Sercan Ö Arık. Chain of agents: Large language models collaborating on long-context tasks. *Advances in Neural Information Processing Systems*, 37:132208–132237, 2024.

Zijing Zhang, Ziyang Chen, Mingxiao Li, Zhaopeng Tu, and Xiaolong Li. Rlvmr: Reinforcement learning with verifiable meta-reasoning rewards for robust long-horizon agents. *arXiv preprint arXiv:2507.22844*, 2025b.

Zhanke Zhou, Xiao Feng, Zhaocheng Zhu, Jiangchao Yao, Sanmi Koyejo, and Bo Han. From passive to active reasoning: Can large language models ask the right questions under incomplete information? *arXiv preprint arXiv:2506.08295*, 2025.

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Daisy Zhe Wang, Zhenhailong Wang, Cheng Qian, Robert Tang, Heng Ji, et al. Multiagentbench: Evaluating the collaboration and competition of llm agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8580–8622, 2025.

Deyu Zou, Yongqiang Chen, Jianxiang Wang, Garry Yang, Mufei Li, Qing Da, James Cheng, Pan Li, and Yu Gong. Reducing belief deviation in reinforcement learning for active reasoning of llm agents. In *International Conference on Learning Representations*, volume 2026, pp. 21511–21545, 2026.

A Theoretical Setup, Definitions, and Assumptions

This appendix provides the formal analysis supporting the main system-level results in Sections 4 and 5. It first collects the shared-memory system model, definitions, and assumptions. It then establishes the collaborative-stagnation results underlying Proposition 4.1, analyzes system-level credit assignment corresponding to Theorem 4.2 and Corollary 5.2, and proves the observable-boundary result corresponding to Proposition 5.1.

We collect the system model, analytical quantities, and assumptions before presenting the theoretical results. The assumptions for collaborative stagnation, credit assignment, and observable detection are invoked separately in the results that follow.

A.1 System Dynamics and Shared-Memory Beliefs

We consider an active reasoning environment

$$\mathcal{E} = (S, \{\mathcal{U}_i\}_{i=1}^n, \mathcal{A}^{\mathrm{env}}, \Omega, \mathcal{O}, R, \gamma),$$

where S is the latent state space, $s^\star \in S$ is the fixed ground-truth state within an episode, $\mathcal{U}_i$ is the role-proposal space of agent A_i, $\mathcal{A}^{\mathrm{env}}$ is the environment-facing action space, Ω is the observation space, $\mathcal{O} : S \times \mathcal{A}^{\mathrm{env}} \to \Delta(\Omega)$ is the observation kernel, R is the reward function, and $\gamma \in (0, 1]$ is the discount factor. Here, $\mathcal{O}(\cdot \mid s, a)$ denotes the conditional probability measure over observations given latent state s and environment-facing action a. The formal analysis below assumes that S and Ω are finite or countably discrete, so beliefs and observation laws are probability mass functions and all Bayesian normalizers are written as sums.

Scope of the abstraction. The discrete formulation is used only for the formal belief-space analysis. For tasks with continuous or structured latent variables, S represents a discrete hypothesis abstraction, such as a finite-resolution hypothesis cover. The practical detector need not operate on this abstraction and may instead use a continuous observable proxy. Whenever a uniform observation lower bound is invoked, we restrict attention to finite reachable observation support.

The multi-agent system consists of agents $\mathrm{Agt} = \{A_1, \ldots, A_n\}$ with parameters $\Theta = (\theta_1, \ldots, \theta_n)$. We write the system state at the beginning of round t as

$$X_t = (B_t, b_t^{\mathrm{mem}}, M_t), \qquad B_t = (b_{1,t}, \ldots, b_{n,t}),$$

where $b_{i,t} \in \Delta(S)$ is the belief of agent A_i, $b_t^{\mathrm{mem}} \in \Delta(S)$ is the belief induced by shared memory, and $M_t \in \mathbb{M}$ is the current shared-memory content, with $\mathbb{M}$ denoting the shared-memory content space. Shared memory carries information across rounds and is accessible to later agents, but does not require a specific storage architecture. When S is countably infinite, $\Delta(S)$ denotes probability mass functions with countable support. Define the belief projection and its state space by

$$\bar{X}_t := \mathrm{proj}_{\mathrm{bel}}(X_t) = (B_t, b_t^{\mathrm{mem}}), \qquad \bar{\mathcal{X}}_{\mathrm{sys}} := \Delta(S)^n \times \Delta(S),$$

and the full operational state space by $\mathcal{X}_{\mathrm{sys}} := \bar{\mathcal{X}}_{\mathrm{sys}} \times \mathbb{M}$. The shared-memory content M_t belongs to the operational system and observable detector layer; the oracle construction below is defined only on $\bar{\mathcal{X}}_{\mathrm{sys}}$.

Throughout the analysis, let $\mathcal{F}_t^{\mathrm{act}}$ denote the actual-trajectory pre-action sigma-field at round t, generated only by the operational interaction history up to the current system state:

$$\mathcal{F}_t^{\mathrm{act}} := \sigma\left(X_1, U_1, a_1^{\mathrm{env}}, o_1, \ldots, X_{t-1}, U_{t-1}, a_{t-1}^{\mathrm{env}}, o_{t-1}, X_t\right).$$

Thus X_t is $\mathcal{F}_t^{\mathrm{act}}$-measurable, while the proposal-generation randomness, shared-memory selection randomness, induced environment-facing action, and observation at round t are marginalized out in conditional probabilities and conditional expectations. Conditional on $\mathcal{F}_t^{\mathrm{act}}$, the roles generate an ordered proposal sequence according to the fixed role order:

$$u_{i,t} \sim \pi_{\theta_i}^{\mathrm{prop}}(\cdot \mid X_t, u_{<i,t}), \qquad u_{<i,t} := (u_{1,t}, \ldots, u_{i-1,t}), \qquad U_t = (u_{1,t}, \ldots, u_{n,t}).$$

Here $u_{<1,t}$ is empty. As in the main text, $\pi_{\theta_i}^{\mathrm{prop}}$ is the semantic-proposal law induced by the role's autoregressive token policy $\pi_{\theta_i}^{\mathrm{tok}}$. Conditioning on X_t is shorthand for the policy kernel induced by

the actor-observable shared-memory history; it does not grant the roles direct access to latent beliefs. The corresponding sequential proposal law is

$$\pi_{\Theta}^{\mathrm{prop}}(dU \mid X) := \prod_{i=1}^{n} \pi_{\theta_i}^{\mathrm{prop}}(du_i \mid X, u_{<i}).$$

The fixed shared-memory mechanism maps the current state and ordered proposal sequence (X_t, U_t) to an admissible environment-facing action through the kernel $a_t^{\mathrm{env}} \sim \kappa_{\mathcal{M}}(\cdot \mid X_t, U_t)$, and the environment returns an observation $o_t \sim \mathcal{O}(\cdot \mid s^\star, a_t^{\mathrm{env}})$. Here, $\mathcal{M}$ denotes the fixed, non-trainable shared-memory mechanism, whereas M_t denotes the time-varying shared-memory content. The kernel $\kappa_{\mathcal{M}}$ is not an additional centralized actor; a deterministic action mapping is represented by a Dirac kernel. After marginalizing over the sequential proposal-generation randomness, the induced effective environment-facing policy is

$$\pi_{\Theta,\mathcal{M}}^{\mathrm{env}}(da \mid X) := \int \kappa_{\mathcal{M}}(da \mid X, U)\, \pi_{\Theta}^{\mathrm{prop}}(dU \mid X).$$

The actual system update is $X_{t+1} = \mathfrak{B}_{\Theta}(X_t, U_t, a_t^{\mathrm{env}}, o_t)$.

A.2 System Potential, Oracle, and Stagnation Conditions

Definition A.1 (Agent-, memory-, and system-level potentials). *For agent i with belief $b_{i,t} \in \Delta(S)$ and ground-truth state $s^\star$, define the* agent-level ground-truth log-loss potential *by*

$$\Psi_{i,t} := -\log b_{i,t}(s^\star).$$

For the memory-induced belief $b_t^{\mathrm{mem}} \in \Delta(S)$, define the memory-level potential *by*

$$\Psi_t^{\mathrm{mem}} := -\log b_t^{\mathrm{mem}}(s^\star).$$

Given weights $\alpha \in [0,1]$, $w_i \geq 0$ with $\sum_{i=1}^n w_i = 1$, define the system-level potential *by*

$$\Psi^{\mathrm{sys}}(\bar{X}_t) := \alpha \Psi_t^{\mathrm{mem}} + (1-\alpha) \sum_{i=1}^{n} w_i \Psi_{i,t}.$$

For a full operational state, set $\Psi^{\mathrm{sys}}(X_t) := \Psi^{\mathrm{sys}}(\bar{X}_t)$. Thus the theoretical potential depends on X_t only through its belief projection, not directly through the shared-memory content M_t. For notational simplicity, write $\Psi_t^{\mathrm{sys}} := \Psi^{\mathrm{sys}}(X_t)$. Assumption A.6 ensures that these logarithms are finite. Lower values indicate higher truth confidence.

Throughout the following actual-system analysis, the conditioning field remains $\mathcal{F}_t^{\mathrm{act}}$. Thus X_t is $\mathcal{F}_t^{\mathrm{act}}$-measurable, while round-t proposal, selection, action, and observation randomness is integrated out. The effective environment-facing policy induced by the role policies and fixed shared-memory mechanism is denoted by $\pi_{\Theta,\mathcal{M}}^{\mathrm{env}}(\cdot \mid X_t)$.

Assumption A.2 (System-level non-degenerate observations). *There exists a constant $\eta > 0$ such that, for every admissible environment-facing action $a \in \mathcal{A}^{\mathrm{env}}$, every state $s \in S$, and every reachable observation $o \in \Omega$, $\mathcal{O}(o \mid s, a) \geq \eta$.*

Before defining system-level oracle informativeness, we specify a *belief-only* analytical oracle. For any $\bar{X} = (B, b^{\mathrm{mem}}) \in \bar{\mathcal{X}}_{\mathrm{sys}}$, with $B = (b_1, \ldots, b_n)$, any environment-facing action a^{env}, and any observation o, the oracle updates each belief component by

$$\mathcal{B}_{\mathrm{Bayes}}^{\star}(b, a^{\mathrm{env}}, o)(s) := \frac{\mathcal{O}(o \mid s, a^{\mathrm{env}}) b(s)}{\sum_{s' \in S} \mathcal{O}(o \mid s', a^{\mathrm{env}}) b(s')}, \qquad s \in S.$$

The denominator is strictly positive on every reachable observation by Assumptions A.6 and A.2. Define $\bar{\mathfrak{B}}^\star(\bar{X}, a^{\mathrm{env}}, o) = (B^+, b^{\mathrm{mem},+})$, with $B^+ = (b_1^+, \ldots, b_n^+)$, by

$$b^{\mathrm{mem},+} = \mathcal{B}_{\mathrm{Bayes}}^{\star}(b^{\mathrm{mem}}, a^{\mathrm{env}}, o), \qquad b_i^+ = \mathcal{B}_{\mathrm{Bayes}}^{\star}(b_i, a^{\mathrm{env}}, o), \qquad i = 1, \ldots, n.$$

No oracle counterpart of the shared-memory content M_t is introduced. On a product probability space, the independently evolving oracle reference samples its action from the abstract belief-only rule $a_t^{\mathrm{env},\star} \sim \pi_{\mathrm{or}}^{\mathrm{env}}\big(\cdot \mid \bar{X}_t^\star\big)$, followed by $o_t^\star \sim \mathcal{O}(\cdot \mid s^\star, a_t^{\mathrm{env},\star})$ and $\bar{X}_{t+1}^\star = \bar{\mathfrak{B}}^\star\big(\bar{X}_t^\star, a_t^{\mathrm{env},\star}, o_t^\star\big)$.

Thus $\bar{X}_t^\star = (B_t^\star, b_t^{\mathrm{mem},\star})$ is an analytical belief trajectory, whereas shared-memory content belongs only to the actual operational trajectory. Conditional on the episode truth and initial states, the oracle innovations are independent of the actual-rollout innovations. All probability laws in this coupled actual–oracle construction are understood conditional on the episode-wise fixed truth $s^\star$ and the initial actual and oracle states. For analyses that simultaneously involve both trajectories, define the joint pre-action filtration

$$\mathcal{G}_t := \sigma\big(\mathcal{F}_t^{\mathrm{act}}, \bar{X}_1^\star, a_1^{\mathrm{env},\star}, o_1^\star, \ldots, a_{t-1}^{\mathrm{env},\star}, o_{t-1}^\star, \bar{X}_t^\star\big)\,.$$

Then X_t and $\bar{X}_t^\star$ are both $\mathcal{G}_t$-measurable. The actual filtration $\mathcal{F}_t^{\mathrm{act}}$ is not enlarged and remains the conditioning field for actual-system progress, CSR, critic, GAE, and detector quantities.

Remark (Counterfactual update versus independent oracle trajectory). The counterfactual one-step oracle update $\bar{\mathfrak{B}}^\star(\bar{X}_t, a_t^{\mathrm{env}}, o_t)$, used below in the one-step actual–oracle potential gap, starts from the actual belief projection and shares the actual action–observation pair. It is distinct from the independently evolving reference state

$$\bar{X}_{t+1}^\star = \bar{\mathfrak{B}}^\star\big(\bar{X}_t^\star, a_t^{\mathrm{env},\star}, o_t^\star\big).$$

In general, these two oracle outputs are not equal.

Definition A.3 (System-level oracle informativeness). *Given a belief state $\bar{X} \in \bar{\mathcal{X}}_{\mathrm{sys}}$ and an environment-facing action $a^{\mathrm{env}} \in \mathcal{A}^{\mathrm{env}}$, define the system-level oracle informativeness by*

$$\mathcal{I}^{\mathrm{sys}}(\bar{X}, a^{\mathrm{env}}) := \Psi^{\mathrm{sys}}(\bar{X}) - \mathbb{E}_{o\sim\mathcal{O}(\cdot|s^\star,a^{\mathrm{env}})}\left[\Psi^{\mathrm{sys}}\left(\bar{\mathfrak{B}}^\star(\bar{X}, a^{\mathrm{env}}, o)\right)\right].$$

This quantity measures the expected potential decrease under the belief-only oracle update. Along the actual trajectory, we abbreviate $\mathcal{I}_t^{\mathrm{sys}} := \mathcal{I}^{\mathrm{sys}}(\bar{X}_t, a_t^{\mathrm{env}})$.

Definition A.4 (Policy-averaged one-step actual–oracle potential gap). *Given the multi-agent parameters Θ and the fixed shared-memory mechanism $\mathcal{M}$, the* policy-averaged one-step actual–oracle potential gap *at system state X_t is defined as*

$$c_{\Theta,\mathcal{M}}^{\mathrm{sys}}(X_t) := \mathbb{E}\left[\Psi^{\mathrm{sys}}\left(\mathfrak{B}_\Theta(X_t, U_t, a_t^{\mathrm{env}}, o_t)\right) - \Psi^{\mathrm{sys}}\left(\bar{\mathfrak{B}}^\star(\bar{X}_t, a_t^{\mathrm{env}}, o_t)\right) \mid \mathcal{F}_t^{\mathrm{act}}\right],$$

where the conditional expectation is taken over the proposal randomness U_t, the shared-memory selection randomness, the induced environment-facing action $a_t^{\mathrm{env}} \sim \pi_{\Theta,\mathcal{M}}^{\mathrm{env}}(\cdot \mid X_t)$, and the environment observation $o_t \sim \mathcal{O}(\cdot \mid s^\star, a_t^{\mathrm{env}})$. Here $\mathfrak{B}_\Theta(X_t, U_t, a_t^{\mathrm{env}}, o_t)$ denotes the updated system state produced by the actual multi-agent update under parameters Θ given the ordered proposal sequence U_t, while $\bar{\mathfrak{B}}^\star(\bar{X}_t, a_t^{\mathrm{env}}, o_t)$ denotes the corresponding updated belief state produced by the analytical oracle. The same realized action and observation are used in this one-step counterfactual comparison; this does not construct an oracle shared-memory-content trajectory. Positive values indicate that the actual collaborative update leaves higher system potential than the corresponding one-step oracle update.

Definition A.5 (System-level one-step progress). *Given the multi-agent parameters Θ and the fixed shared-memory mechanism $\mathcal{M}$, the* policy-averaged system-level expected progress *at system state X_t is defined as*

$$\mathcal{P}_{\Theta,\mathcal{M}}^{\mathrm{sys}}(X_t) := \Psi^{\mathrm{sys}}(X_t) - \mathbb{E}\left[\Psi^{\mathrm{sys}}\left(\mathfrak{B}_\Theta(X_t, U_t, a_t^{\mathrm{env}}, o_t)\right) \mid \mathcal{F}_t^{\mathrm{act}}\right],$$

where the conditional expectation is with respect to the same round-t operational randomness as above. Thus, $\mathcal{P}_{\Theta,\mathcal{M}}^{\mathrm{sys}}(X_t)$ is the policy-averaged expected one-step decrease of the system-level potential.

Since the actual update satisfies $X_{t+1} = \mathfrak{B}_\Theta(X_t, U_t, a_t^{\mathrm{env}}, o_t)$, Definition A.5 immediately implies $\mathbb{E}\left[\Psi_{t+1}^{\mathrm{sys}} \mid \mathcal{F}_t^{\mathrm{act}}\right] = \Psi_t^{\mathrm{sys}} - \mathcal{P}_{\Theta,\mathcal{M}}^{\mathrm{sys}}(X_t)$. Hence, if $\mathcal{P}_{\Theta,\mathcal{M}}^{\mathrm{sys}}(X_t) \le 0$, then the expected system-level potential does not decrease at X_t.

Assumption A.6 (Strict truth support and potential integrability). *The initial actual beliefs and the initial oracle beliefs assign positive mass to $s^\star$. Moreover, the actual LLM-mediated update preserves truth support on the reachable states included in the analysis: for every analyzed round t and role i, almost surely,*

$$b_{i,t}(s^\star) > 0, \qquad b_t^{\mathrm{mem}}(s^\star) > 0, \qquad i = 1, \ldots, n.$$

The resulting actual system potential is integrable at every analyzed round: $\mathbb{E}[\Psi_t^{\mathrm{sys}}] < \infty$. *In addition, the random one-step potential difference underlying* $c^{\mathrm{sys}}_{\Theta,\mathcal{M}}(X_t)$ *is integrable at every analyzed round:*

$$\mathbb{E}\big[\big|\Psi^{\mathrm{sys}}(\mathfrak{B}_\Theta(X_t, U_t, a_t^{\mathrm{env}}, o_t)) - \Psi^{\mathrm{sys}}\big(\bar{\mathfrak{B}}^\star(\bar{X}_t, a_t^{\mathrm{env}}, o_t)\big)\big|\big] < \infty.$$

The corresponding oracle support is preserved by the Bayesian update under Assumption A.2. Consequently, actual potentials are finite almost surely and their expectations are well defined, while Assumption A.7 supplies the separate uniform oracle bound needed by the threshold defining the CSR. No uniform positive lower bound on the actual truth mass is imposed.

In particular, if $b(s^\star) > 0$, then every reachable oracle update preserves truth support:

$$\mathcal{B}^\star_{\mathrm{Bayes}}(b, a, o)(s^\star) = \frac{\mathcal{O}(o \mid s^\star, a) b(s^\star)}{\sum_{s \in S} \mathcal{O}(o \mid s, a) b(s)} > 0.$$

Assumption A.7 (Bounded oracle potential). *There exists a constant* $\Psi_{\mathrm{or}} > 0$ *such that, for all time steps* t *considered in the analysis, the* oracle system-level potential *satisfies*

$$\Psi_t^{\mathrm{sys},\star} := \Psi^{\mathrm{sys}}(\bar{X}_t^\star) \le \Psi_{\mathrm{or}} \qquad \textit{almost surely},$$

where $\bar{X}_t^\star$ *is the belief-only oracle state generated by* $\pi_{\mathrm{or}}^{\mathrm{env}}$ *and* $\bar{\mathfrak{B}}^\star$.

Since $\Psi^{\mathrm{sys}}(\cdot) \ge 0$, Assumption A.7 implies that, for any environment-facing action $a^{\mathrm{env}} \in \mathcal{A}^{\mathrm{env}}$,

$$\mathcal{I}^{\mathrm{sys}}(\bar{X}_t^\star, a^{\mathrm{env}}) = \Psi^{\mathrm{sys}}(\bar{X}_t^\star) - \mathbb{E}_{o \sim \mathcal{O}(\cdot \mid s^\star, a^{\mathrm{env}})}\big[\Psi^{\mathrm{sys}}\big(\bar{\mathfrak{B}}^\star(\bar{X}_t^\star, a^{\mathrm{env}}, o)\big)\big] \le \Psi_{\mathrm{or}}.$$

Consequently, for any action-selection rule used in the oracle analysis,

$$\mathbb{E}\big[\mathcal{I}^{\mathrm{sys}}(\bar{X}_t^\star, a_t^{\mathrm{env},\star}) \,\big|\, \mathcal{G}_t\big] \le \Psi_{\mathrm{or}}.$$

Definition A.8 (Shared-memory-mediated collaborative stagnation region). *Given the multi-agent parameters* Θ *and the fixed, non-trainable shared-memory mechanism* $\mathcal{M}$, *a set* $\mathcal{R}^{\mathrm{sys}}_{\Theta,\mathcal{M}} \subseteq \Delta(S)^n \times \Delta(S) \times \mathbb{M}$ *is called a* collaborative stagnation region (CSR) *if, for every reachable system state* $X_t \in \mathcal{R}^{\mathrm{sys}}_{\Theta,\mathcal{M}}$, *the actual policy-averaged update remains in* $\mathcal{R}^{\mathrm{sys}}_{\Theta,\mathcal{M}}$ *almost surely and induces non-positive expected system-level progress:*

$$\mathbb{P}\left(X_{t+1} \in \mathcal{R}^{\mathrm{sys}}_{\Theta,\mathcal{M}} \mid \mathcal{F}_t^{\mathrm{act}}\right) = 1, \qquad \mathbb{E}\left[\Psi^{\mathrm{sys}}_{t+1} \mid \mathcal{F}_t^{\mathrm{act}}\right] \ge \Psi_t^{\mathrm{sys}}.$$

Here $X_t = (B_t, b_t^{\mathrm{mem}}, M_t)$ *denotes the system state, while* $\Psi_t^{\mathrm{sys}} = \Psi^{\mathrm{sys}}(X_t)$ *denotes the system-level potential.*

Definition A.9 (Belief-projection distance). *For two belief states* $\bar{X} = (B, b^{\mathrm{mem}})$ *and* $\bar{X}' = (B', b^{\mathrm{mem}\prime})$, *where* $B = (b_1, \ldots, b_n)$ *and* $B' = (b'_1, \ldots, b'_n)$, *define*

$$d_{\mathrm{bel}}(\bar{X}, \bar{X}') := \omega_{\mathrm{mem}} \left\|b^{\mathrm{mem}} - b^{\mathrm{mem}\prime}\right\|_1 + \sum_{i=1}^{n} \omega_i \left\|b_i - b'_i\right\|_1,$$

where $\omega_{\mathrm{mem}} = \alpha, \quad \omega_i = (1-\alpha) w_i, \quad i = 1, \ldots, n.$

This is a weighted pseudometric on $\bar{\mathcal{X}}_{\mathrm{sys}}$, *and it is a metric when all component weights are strictly positive. In either case,* $0 \le d_{\mathrm{bel}}(\bar{X}, \bar{X}') \le 2$. *It is not used as a metric on shared-memory content and therefore does not, by itself, constrain the operational policy* $\pi^{\mathrm{env}}_{\Theta,\mathcal{M}}(\cdot \mid X)$.

Assumption A.10 (System-level one-step actual–oracle gap growth). *There exist constants* $m_{\Theta,\mathcal{M}} > 0$, $c_0^{\mathrm{sys}} \ge 0$, *and* $\Psi_0^{\mathrm{sys}} \ge 0$ *such that, for every reachable system state* $X \in \mathcal{X}_{\mathrm{sys}}$, *if* $\Psi^{\mathrm{sys}}(X) \ge \Psi_0^{\mathrm{sys}}$, *then the policy-averaged one-step actual–oracle potential gap satisfies*

$$c^{\mathrm{sys}}_{\Theta,\mathcal{M}}(X) \ge m_{\Theta,\mathcal{M}} \Psi^{\mathrm{sys}}(X) - c_0^{\mathrm{sys}}.$$

Equivalently, along the actual multi-agent trajectory $(X_t)_{t \ge 1}$, *whenever* $\Psi_t^{\mathrm{sys}} := \Psi^{\mathrm{sys}}(X_t) \ge \Psi_0^{\mathrm{sys}}$, *we have*

$$c^{\mathrm{sys}}_{\Theta,\mathcal{M}}(X_t) \ge m_{\Theta,\mathcal{M}} \Psi_t^{\mathrm{sys}} - c_0^{\mathrm{sys}}.$$

Here, $c^{\mathrm{sys}}_{\Theta,\mathcal{M}}(X)$ *denotes the policy-averaged one-step actual–oracle potential gap between the actual shared-memory-mediated update and the belief-only oracle update applied to* $\mathrm{proj}_{\mathrm{bel}}(X)$. *In reachable high-potential states, the one-step actual–oracle potential gap grows at least linearly with the system-level potential.*

Assumption A.11 (Bounded policy-averaged oracle-informativeness mismatch). *There exists a finite constant $C_{\mathrm{mis}} \geq 0$ such that, almost surely for every analyzed round t,*

$$\left| \mathbb{E}_{a^{\mathrm{env}} \sim \pi_{\mathrm{or}}^{\mathrm{env}}(\cdot|\bar{X}_t^\star)} \mathcal{I}^{\mathrm{sys}}(\bar{X}_t^\star, a^{\mathrm{env}}) - \mathbb{E}_{a^{\mathrm{env}} \sim \pi_{\Theta,\mathcal{M}}^{\mathrm{env}}(\cdot|X_t)} \mathcal{I}^{\mathrm{sys}}(\bar{X}_t, a^{\mathrm{env}}) \right| \leq C_{\mathrm{mis}}.$$

By Lemma B.2, $\mathcal{I}^{\mathrm{sys}} \in [0, \Lambda_I]$, so the mismatch is always finite and one universal admissible choice is $C_{\mathrm{mis}} = \Lambda_I = -\log \eta$. We retain C_{mis} explicitly to allow a sharper problem-dependent mismatch bound.

Remark A.12. ***(Optional belief-only sufficient condition).*** *Lemmas B.1 and B.2 may yield a tighter value of C_{mis} when both action rules factor through belief projections and satisfy an appropriate total-variation Lipschitz condition. The actual shared-memory implementation need not satisfy this factorization because its actors may read M_t directly, and the main CSR results do not depend on this additional factorization.*

Throughout the following results, actual-only conditional quantities use $\mathcal{F}_t^{\mathrm{act}}$. Conditional expectations that simultaneously contain an actual quantity and the independently evolving oracle reference use $\mathcal{G}_t$. Hence $X_t, \bar{X}_t, \bar{X}_t^\star$ are $\mathcal{G}_t$-measurable, while the corresponding round-t innovations are integrated out.

For the remaining results, fix a finite rollout horizon with terminal state index T. With the constants introduced above, write $K_{\mathrm{sys}} := \Psi_{\mathrm{or}} + C_{\mathrm{mis}}$ and

$$\begin{aligned} \Psi_{\mathrm{CSR}} &:= \max\left\{ \Psi_0^{\mathrm{sys}}, \frac{K_{\mathrm{sys}} + c_0^{\mathrm{sys}}}{m_{\Theta,\mathcal{M}}} \right\}, \\ \mathcal{R}_{\Theta,\mathcal{M}}^{\mathrm{sys}}(\Psi_{\mathrm{CSR}}) &:= \{ X \in \mathcal{X}_{\mathrm{sys}} : \Psi^{\mathrm{sys}}(X) \geq \Psi_{\mathrm{CSR}} \}, \\ t_S^{\mathrm{sys}} &:= \inf\{ t \geq 1 : \Psi_t^{\mathrm{sys}} \geq \Psi_{\mathrm{CSR}} \}, \qquad \mathsf{E}_\tau := \{ t_S^{\mathrm{sys}} = \tau \}. \end{aligned}$$

The collaborative-stagnation results below establish conditions under which this candidate threshold region is a CSR. The credit analysis additionally assumes post-entry confidence decline; that assumption is separate from the CSR result.

A.3 Observable Progress and the Practical Boundary

Along the actual shared-memory-mediated multi-agent trajectory $(X_t)_{t \geq 1}$, define the *realized* one-step system potential progress by

$$G_t^{\mathrm{sys}} := \Psi_t^{\mathrm{sys}} - \Psi_{t+1}^{\mathrm{sys}} = \Psi^{\mathrm{sys}}(X_t) - \Psi^{\mathrm{sys}}(X_{t+1}).$$

This quantity is distinct from the policy-averaged one-step progress $\mathcal{P}_{\Theta,\mathcal{M}}^{\mathrm{sys}}(X_t)$ defined earlier. Specifically, G_t^{sys} is the realized post-update drop in the system-level potential, while $\mathcal{P}_{\Theta,\mathcal{M}}^{\mathrm{sys}}(X_t)$ is its conditional expectation given $\mathcal{F}_t^{\mathrm{act}}$.

Let $\mathcal{H}_t^{\mathrm{sys}}$ denote the observable system trace available at the beginning of round t, up to the current state X_t. It contains the current shared-memory content and all role outputs, environment-facing actions, and environment observations from rounds before t. The trace $\mathcal{H}_{t+1}^{\mathrm{sys}}$ additionally contains the observable information produced during round t. It excludes the latent belief components used only for analysis. Let $D_t^{\mathrm{sys}} := D_{\mathrm{obs}}\left(\mathcal{H}_t^{\mathrm{sys}}, \mathcal{H}_{t+1}^{\mathrm{sys}}\right)$ be an $\mathcal{F}_{t+1}^{\mathrm{act}}$-measurable observable proxy signal intended to track G_t^{sys}. Thus, the detector uses the observable trace produced by the transition from round t to round $t+1$, rather than the complete theoretical states X_t and X_{t+1}. Fix a threshold $\Delta_{\min} \in \mathbb{R}$ and a window length $k \in \mathbb{N}$. For any starting index t satisfying $1 \leq t \leq T - k$, define the system-level window-stall event by

$$\mathcal{W}_{t,k}^{\mathrm{sys}} := \bigcap_{j=0}^{k-1} \left\{ D_{t+j}^{\mathrm{sys}} \leq \Delta_{\min} \right\}.$$

This is the event that all k observable system-level proxy values in the window are classified as low progress. Define the practical observable optimization boundary by

$$\widehat{H} := \inf\left(\left\{ t + k : 1 \leq t \leq T - k,\ \mathcal{W}_{t,k}^{\mathrm{sys}} \text{ holds} \right\} \cup \{T\} \right).$$

If no window triggers, then $\widehat{H} = T$. Otherwise, the boundary is placed at the end of the first triggering window: the confirmation window and $X_{\widehat{H}}$ are retained, and only the post-confirmation continuation is excluded from optimization. Thus $\widehat{H}$ is the practical observable boundary; it is not claimed to equal, recover, or consistently estimate the latent CSR entry time t_S^{sys}. Proposition D.2 provides neither a bound on $|\widehat{H} - t_S^{\mathrm{sys}}|$ nor a detection-delay guarantee.

A.4 Critic, Temporal Credit, and Post-Entry Conditions

Definition A.13 (System truth-confidence induced by the system potential). *For any system state $X \in \mathcal{X}_{\mathrm{sys}}$ with belief projection $\mathrm{proj}_{\mathrm{bel}}(X) = (B, b^{\mathrm{mem}})$, where $B = (b_1, \ldots, b_n)$, define the system truth-confidence by*

$$\chi^{\mathrm{sys}}(X) := \exp(-\Psi^{\mathrm{sys}}(X)) = (b^{\mathrm{mem}}(s^\star))^{\alpha} \prod_{i=1}^{n} (b_i(s^\star))^{(1-\alpha)w_i}. \tag{1}$$

Along the actual trajectory, write $\chi_t^{\mathrm{sys}} := \chi^{\mathrm{sys}}(X_t)$. Thus $\chi_t^{\mathrm{sys}} \in (0, 1]$ under Assumption A.6, and $\Psi_t^{\mathrm{sys}} = -\log \chi_t^{\mathrm{sys}}$. Hence lower system-level potential is equivalent to higher system truth-confidence.

Definition A.14 (System-level TD residual and GAE). *Fix a finite rollout horizon with transition rounds $1, \ldots, T-1$ and state X_T. Let $V_t^{\mathrm{sys}} := V^{\mathrm{sys}}(X_t)$ be the centralized system-level critic used only during training, and set $V_T^{\mathrm{sys}} = 0$. Here X_T denotes a genuine environment terminal state, for which $V_T^{\mathrm{sys}} = 0$. An artificial* PHANTOM-CONTROL *boundary $H < T$ is not treated as terminal. Let r_t denote the realized environment reward associated with transition t. The system-level TD residual is*

$$\delta_t^{\mathrm{sys}} := r_t + \gamma V_{t+1}^{\mathrm{sys}} - V_t^{\mathrm{sys}}, \qquad t = 1, \ldots, T-1. \tag{2}$$

For $\lambda \in [0, 1]$, the system-level generalized advantage estimator assigned to the round-t role-specific proposal or token factors is

$$\widehat{A}_t^{\mathrm{sys}} := \sum_{j=0}^{T-t-1} (\gamma\lambda)^j \delta_{t+j}^{\mathrm{sys}}. \tag{3}$$

For an optimization boundary $H \le T$, define the prefix-supported estimator by

$$\widehat{A}_{t,<H}^{\mathrm{sys}} := \begin{cases} \displaystyle\sum_{\ell=t}^{H-1} (\gamma\lambda)^{\ell-t} \delta_\ell^{\mathrm{sys}}, & t < H, \\ 0, & t \ge H, \end{cases} \qquad t = 1, \ldots, T-1.$$

The zero extension for $t \ge H$ is a bookkeeping convention that makes the masked policy objective well defined; only $t < H$ contributes to optimization. The choice $H = T$ recovers the full-horizon estimator; conditional on E_τ, the ideal boundary is $H = \tau$, whereas the practical boundary is $H = \widehat{H}$. For an artificial boundary $H < T$, the final retained residual remains

$$\delta_{H-1}^{\mathrm{sys}} = r_{H-1} + \gamma V^{\mathrm{sys}}(X_H) - V^{\mathrm{sys}}(X_{H-1}),$$

so $V^{\mathrm{sys}}(X_H)$ is retained for bootstrapping. It is set to zero only when X_H is a genuine environment terminal state. In a shared-reward multi-agent actor-critic implementation, this scalar is the common credit signal applied to the role-specific proposal or token log-likelihood scores at round t.

Assumption A.15 (Outcome-based rewards and terminal residual). *The rollout is outcome-based: $r_\ell = 0$ for all $\ell < T - 1$, and the only nonzero environment reward may appear at the terminal transition $T - 1$. Moreover,*

$$\mathbb{E}\big[\delta_{T-1}^{\mathrm{sys}} \,\big|\, \mathcal{F}_{T-1}^{\mathrm{act}}\big] \le 0. \tag{4}$$

This condition is satisfied, for example, on failed rollouts with terminal reward $r_{T-1} = 0$ and $V_{T-1}^{\mathrm{sys}} \ge 0$. It may be replaced by equality when the terminal critic is calibrated.

Assumption A.16 (System-level value calibration). *There exists an increasing differentiable function $g : [0, 1] \to [0, 1]$ such that, for every nonterminal state along the analyzed trajectory,*

$$V_t^{\mathrm{sys}} = g(\chi_t^{\mathrm{sys}}), \qquad t = 1, \ldots, T-1, \qquad \inf_{x \in [0,1]} g'(x) \ge \kappa_V > 0. \tag{5}$$

The terminal bootstrap convention $V_T^{\mathrm{sys}} = 0$ is separate and is not required to satisfy this calibration relation.

Assumption A.17 (Post-entry system truth-confidence drop). *Assume that the threshold region $\mathcal{R}^{\mathrm{sys}}_{\Theta,\mathcal{M}}(\Psi_{\mathrm{CSR}})$ is absorbing, and suppose the analyzed rollout has entered it. There exists $\rho^{\mathrm{sys}}_{\chi} > 0$ such that, for every ℓ with $t^{\mathrm{sys}}_{S} \le \ell \le T-2$,*

$$\chi^{\mathrm{sys}}_{\ell+1} \le \chi^{\mathrm{sys}}_{\ell} \qquad a.s., \tag{6}$$

and

$$\mathbb{E}\big[\chi^{\mathrm{sys}}_{\ell+1} - \chi^{\mathrm{sys}}_{\ell} \,\big|\, \mathcal{F}^{\mathrm{act}}_{\ell}\big] \le -\rho^{\mathrm{sys}}_{\chi}. \tag{7}$$

B Collaborative-Stagnation Analysis

Lemma B.1 (Belief-state Lipschitz continuity of informativeness). *Under Assumption A.2, for any fixed environment-facing action $a^{\mathrm{env}} \in \mathcal{A}^{\mathrm{env}}$ and any belief states $\bar{X}, \bar{X}' \in \bar{\mathcal{X}}_{\mathrm{sys}}$, we have*

$$\big|\mathcal{I}^{\mathrm{sys}}(\bar{X}, a^{\mathrm{env}}) - \mathcal{I}^{\mathrm{sys}}(\bar{X}', a^{\mathrm{env}})\big| \le \frac{1}{\eta} d_{\mathrm{bel}}(\bar{X}, \bar{X}').$$

Consequently, for any action distribution $q \in \Delta(\mathcal{A}^{\mathrm{env}})$,

$$\big|\mathbb{E}_{a^{\mathrm{env}}\sim q}\mathcal{I}^{\mathrm{sys}}(\bar{X}, a^{\mathrm{env}}) - \mathbb{E}_{a^{\mathrm{env}}\sim q}\mathcal{I}^{\mathrm{sys}}(\bar{X}', a^{\mathrm{env}})\big| \le \frac{1}{\eta} d_{\mathrm{bel}}(\bar{X}, \bar{X}').$$

Proof. For each belief component, we first derive the closed form of the one-step oracle informativeness. Let $b \in \Delta(S)$ be any belief component, which can be either an agent belief b_i or the memory belief b^{mem}. For a fixed environment-facing action a^{env}, define

$$P(o) := \mathcal{O}(o \mid s^\star, a^{\mathrm{env}}), \qquad Q_b(o) := p_b(o \mid a^{\mathrm{env}}) = \sum_{s\in S} b(s)\mathcal{O}(o \mid s, a^{\mathrm{env}}).$$

Let $\mathcal{B}^\star_{\mathrm{Bayes}}$ denote the component-wise Bayesian update induced by $\bar{\mathfrak{B}}^\star$. By definition,

$$\mathcal{B}^\star_{\mathrm{Bayes}}(b, a^{\mathrm{env}}, o)(s^\star) = \frac{\mathcal{O}(o \mid s^\star, a^{\mathrm{env}})b(s^\star)}{p_b(o \mid a^{\mathrm{env}})}.$$

Therefore,

$$\begin{aligned}
\mathcal{I}(b, a^{\mathrm{env}}) &= \Psi(b) - \mathbb{E}_{o\sim\mathcal{O}(\cdot\mid s^\star, a^{\mathrm{env}})}\left[\Psi\big(\mathcal{B}^\star_{\mathrm{Bayes}}(b, a^{\mathrm{env}}, o)\big)\right] \\
&= -\log b(s^\star) - \mathbb{E}_{o\sim P}\left[-\log\left(\frac{\mathcal{O}(o \mid s^\star, a^{\mathrm{env}})b(s^\star)}{p_b(o \mid a^{\mathrm{env}})}\right)\right] \\
&= \mathbb{E}_{o\sim P}\left[\log\frac{P(o)}{Q_b(o)}\right].
\end{aligned}$$

By Assumption A.2, for every reachable observation o, $Q_b(o) = \sum_{s\in S} b(s)\mathcal{O}(o \mid s, a^{\mathrm{env}}) \ge \sum_{s\in S} b(s)\eta = \eta$. Similarly, $Q_{b'}(o) \ge \eta$. For any $x, y \ge \eta > 0$, we have $|\log x - \log y| \le \frac{|x-y|}{\eta}$. Applying this inequality to $Q_b(o)$ and $Q_{b'}(o)$, we obtain

$$|\log Q_b(o) - \log Q_{b'}(o)| \le \frac{|Q_b(o) - Q_{b'}(o)|}{\eta}.$$

Since $\mathcal{I}(b, a^{\mathrm{env}}) = \mathbb{E}_{o\sim P}[\log P(o)] - \mathbb{E}_{o\sim P}[\log Q_b(o)]$, where the first term is constant in b, we have

$$\begin{aligned}
|\mathcal{I}(b, a^{\mathrm{env}}) - \mathcal{I}(b', a^{\mathrm{env}})| &\le \mathbb{E}_{o\sim P}\left[|\log Q_b(o) - \log Q_{b'}(o)|\right] \\
&\le \frac{1}{\eta}\mathbb{E}_{o\sim P}\left[|Q_b(o) - Q_{b'}(o)|\right] \\
&\le \frac{1}{\eta}\|Q_b - Q_{b'}\|_1.
\end{aligned}$$

Moreover,

$$\begin{aligned}\|Q_b - Q_{b'}\|_1 &= \sum_{o\in\Omega}\left|\sum_{s\in S}(b(s)-b'(s))\mathcal{O}(o \mid s, a^{\mathrm{env}})\right| \\ &\le \sum_{o\in\Omega}\sum_{s\in S}|b(s)-b'(s)|\mathcal{O}(o \mid s, a^{\mathrm{env}}) \\ &= \sum_{s\in S}|b(s)-b'(s)|\sum_{o\in\Omega}\mathcal{O}(o \mid s, a^{\mathrm{env}}) \\ &= \|b-b'\|_1.\end{aligned}$$

Thus, for any belief component, $|\mathcal{I}(b, a^{\mathrm{env}}) - \mathcal{I}(b', a^{\mathrm{env}})| \le \frac{1}{\eta}\|b-b'\|_1$.

Now consider the system-level informativeness. Let $\bar{X} = (B, b^{\mathrm{mem}})$ and $\bar{X}' = (B', b^{\mathrm{mem}\prime})$, where $B = (b_1, \ldots, b_n)$ and $B' = (b'_1, \ldots, b'_n)$. By the component-wise definition of the oracle system update and the additive definition of Ψ^{sys}, we can write

$$\mathcal{I}^{\mathrm{sys}}(\bar{X}, a^{\mathrm{env}}) = \omega_{\mathrm{mem}}\mathcal{I}(b^{\mathrm{mem}}, a^{\mathrm{env}}) + \sum_{i=1}^{n}\omega_i\mathcal{I}(b_i, a^{\mathrm{env}}),$$

where $\omega_{\mathrm{mem}} = \alpha, \omega_i = (1-\alpha)w_i$. Therefore,

$$\begin{aligned}&\left|\mathcal{I}^{\mathrm{sys}}(\bar{X}, a^{\mathrm{env}}) - \mathcal{I}^{\mathrm{sys}}(\bar{X}', a^{\mathrm{env}})\right| \\ \le\ & \omega_{\mathrm{mem}}\left|\mathcal{I}(b^{\mathrm{mem}}, a^{\mathrm{env}}) - \mathcal{I}(b^{\mathrm{mem}\prime}, a^{\mathrm{env}})\right| + \sum_{i=1}^{n}\omega_i\left|\mathcal{I}(b_i, a^{\mathrm{env}}) - \mathcal{I}(b'_i, a^{\mathrm{env}})\right| \\ \le\ & \frac{1}{\eta}\left(\omega_{\mathrm{mem}}\|b^{\mathrm{mem}} - b^{\mathrm{mem}\prime}\|_1 + \sum_{i=1}^{n}\omega_i\|b_i - b'_i\|_1\right) \\ \le\ & \frac{1}{\eta}d_{\mathrm{bel}}(\bar{X}, \bar{X}').\end{aligned}$$

This proves the pointwise system-state Lipschitz bound. Finally, for any action distribution $q \in \Delta(\mathcal{A}^{\mathrm{env}})$, we have

$$\begin{aligned}&\left|\mathbb{E}_{a^{\mathrm{env}}\sim q}\mathcal{I}^{\mathrm{sys}}(\bar{X}, a^{\mathrm{env}}) - \mathbb{E}_{a^{\mathrm{env}}\sim q}\mathcal{I}^{\mathrm{sys}}(\bar{X}', a^{\mathrm{env}})\right| \\ =\ & \left|\mathbb{E}_{a^{\mathrm{env}}\sim q}\left[\mathcal{I}^{\mathrm{sys}}(\bar{X}, a^{\mathrm{env}}) - \mathcal{I}^{\mathrm{sys}}(\bar{X}', a^{\mathrm{env}})\right]\right| \\ \le\ & \mathbb{E}_{a^{\mathrm{env}}\sim q}\left[\left|\mathcal{I}^{\mathrm{sys}}(\bar{X}, a^{\mathrm{env}}) - \mathcal{I}^{\mathrm{sys}}(\bar{X}', a^{\mathrm{env}})\right|\right] \\ \le\ & \frac{1}{\eta}d_{\mathrm{bel}}(\bar{X}, \bar{X}').\end{aligned}$$

This completes the proof. □

Lemma B.2 (Action-distribution continuity of system-level informativeness). *Under Assumption A.2, for any fixed belief state $\bar{X} \in \bar{\mathcal{X}}_{\mathrm{sys}}$ and any two environment-facing action distributions $q, q' \in \Delta(\mathcal{A}^{\mathrm{env}})$, we have*

$$\left|\mathbb{E}_{a^{\mathrm{env}}\sim q}\mathcal{I}^{\mathrm{sys}}(\bar{X}, a^{\mathrm{env}}) - \mathbb{E}_{a^{\mathrm{env}}\sim q'}\mathcal{I}^{\mathrm{sys}}(\bar{X}, a^{\mathrm{env}})\right| \le \Lambda_I d_{\mathrm{TV}}(q, q'),$$

where $\Lambda_I := -\log\eta$ and $d_{\mathrm{TV}}(P, Q) := \sup_{A\subseteq\mathcal{A}^{\mathrm{env}}}|P(A) - Q(A)|$.

Proof. Fix a belief state $\bar{X} = (B, b^{\mathrm{mem}})$ with $B = (b_1, \ldots, b_n)$. For any belief component $b \in \Delta(S)$ and any environment-facing action $a^{\mathrm{env}} \in \mathcal{A}^{\mathrm{env}}$, define

$$P_{a^{\mathrm{env}}}(o) := \mathcal{O}(o \mid s^\star, a^{\mathrm{env}}), \qquad Q_{b,a^{\mathrm{env}}}(o) := p_b(o \mid a^{\mathrm{env}}) = \sum_{s\in S}b(s)\mathcal{O}(o \mid s, a^{\mathrm{env}}).$$

By the component-wise oracle Bayesian update, the one-step oracle informativeness of the belief component b can be written as

$$\begin{aligned}\mathcal{I}(b, a^{\mathrm{env}}) &= \mathbb{E}_{o\sim\mathcal{O}(\cdot\mid s^\star, a^{\mathrm{env}})}\left[\log\frac{\mathcal{O}(o \mid s^\star, a^{\mathrm{env}})}{p_b(o \mid a^{\mathrm{env}})}\right] \\ &= D_{\mathrm{KL}}\big(P_{a^{\mathrm{env}}} \,\|\, Q_{b,a^{\mathrm{env}}}\big).\end{aligned}$$

Hence $\mathcal{I}(b, a^{\mathrm{env}}) \geq 0$. Moreover, by Assumption A.2, for every reachable observation o,

$$Q_{b,a^{\mathrm{env}}}(o) = \sum_{s \in S} b(s)\mathcal{O}(o \mid s, a^{\mathrm{env}}) \geq \sum_{s \in S} b(s)\eta = \eta.$$

Since $P_{a^{\mathrm{env}}}(o) \leq 1$, we have $\log \frac{P_{a^{\mathrm{env}}}(o)}{Q_{b,a^{\mathrm{env}}}(o)} \leq \log \frac{1}{\eta} = -\log \eta$. Therefore, $0 \leq \mathcal{I}(b, a^{\mathrm{env}}) \leq -\log \eta$.

Now consider the system-level oracle informativeness. By the component-wise definition of the oracle system update and the additive definition of Ψ^{sys}, we have

$$\mathcal{I}^{\mathrm{sys}}(\bar{X}, a^{\mathrm{env}}) = \omega_{\mathrm{mem}}\mathcal{I}(b^{\mathrm{mem}}, a^{\mathrm{env}}) + \sum_{i=1}^{n} \omega_i \mathcal{I}(b_i, a^{\mathrm{env}}),$$

where $\omega_{\mathrm{mem}} = \alpha, \quad \omega_i = (1-\alpha)w_i, \quad i = 1, \ldots, n$. Since $\omega_{\mathrm{mem}} \geq 0, \omega_i \geq 0$, and

$$\omega_{\mathrm{mem}} + \sum_{i=1}^{n} \omega_i = \alpha + (1-\alpha)\sum_{i=1}^{n} w_i = 1,$$

the component-wise bound implies $0 \leq \mathcal{I}^{\mathrm{sys}}(\bar{X}, a^{\mathrm{env}}) \leq -\log \eta = \Lambda_I$. Define $f_{\bar{X}}(a^{\mathrm{env}}) := \mathcal{I}^{\mathrm{sys}}(\bar{X}, a^{\mathrm{env}})$. Then $0 \leq f_{\bar{X}}(a^{\mathrm{env}}) \leq \Lambda_I$ for all admissible environment-facing actions. By the variational characterization of total variation, for any measurable function f satisfying $0 \leq f \leq \Lambda_I$, we have

$$|\mathbb{E}_{a \sim q} f(a) - \mathbb{E}_{a \sim q'} f(a)| \leq \Lambda_I d_{\mathrm{TV}}(q, q').$$

Applying this inequality to $f = f_{\bar{X}}$ gives

$$\left|\mathbb{E}_{a^{\mathrm{env}} \sim q}\mathcal{I}^{\mathrm{sys}}(\bar{X}, a^{\mathrm{env}}) - \mathbb{E}_{a^{\mathrm{env}} \sim q'}\mathcal{I}^{\mathrm{sys}}(\bar{X}, a^{\mathrm{env}})\right| \leq \Lambda_I d_{\mathrm{TV}}(q, q').$$

This completes the proof. □

Proposition B.3 (Non-positive system-level drift above the CSR threshold). *Suppose Assumptions A.6, A.7, A.2, A.10, and A.11 hold in the policy-averaged sense described above. Define* $K_{\mathrm{sys}} := \Psi_{\mathrm{or}} + C_{\mathrm{mis}}$*, and the threshold* $\Psi_{\mathrm{CSR}} := \max\left\{\Psi_0^{\mathrm{sys}}, \frac{K_{\mathrm{sys}} + c_0^{\mathrm{sys}}}{m_{\Theta,\mathcal{M}}}\right\}$*. If, at some time step t, the actual system-level potential satisfies* $\Psi_t^{\mathrm{sys}} \geq \Psi_{\mathrm{CSR}}$*, then* $\mathcal{P}^{\mathrm{sys}}_{\Theta,\mathcal{M}}(X_t) \leq 0$*, and hence* $\mathbb{E}\left[\Psi_{t+1}^{\mathrm{sys}} \mid \mathcal{F}_t^{\mathrm{act}}\right] \geq \Psi_t^{\mathrm{sys}}$*. Consequently, if the threshold region*

$$\mathcal{R}^{\mathrm{sys}}_{\Theta,\mathcal{M}}(\Psi_{\mathrm{CSR}}) := \{X \in \mathcal{X}_{\mathrm{sys}} : \Psi^{\mathrm{sys}}(X) \geq \Psi_{\mathrm{CSR}}\}$$

is absorbing under the actual multi-agent update, and if the above statewise conditions hold for every reachable $X \in \mathcal{R}^{\mathrm{sys}}_{\Theta,\mathcal{M}}(\Psi_{\mathrm{CSR}})$*, then it is a shared-memory-mediated collaborative stagnation region.*

Proof. Let $\pi_t := \pi^{\mathrm{env}}_{\Theta,\mathcal{M}}(\cdot \mid X_t), \quad \pi_t^\star := \pi^{\mathrm{env}}_{\mathrm{or}}(\cdot \mid \bar{X}_t^\star)$. By the policy-averaged definitions of system-level oracle informativeness, the one-step actual–oracle potential gap, and system-level one-step progress, we have

$$\mathcal{P}^{\mathrm{sys}}_{\Theta,\mathcal{M}}(X_t) = \mathbb{E}_{a^{\mathrm{env}} \sim \pi_t}\left[\mathcal{I}^{\mathrm{sys}}(\bar{X}_t, a^{\mathrm{env}})\right] - c^{\mathrm{sys}}_{\Theta,\mathcal{M}}(X_t).$$

We first upper bound the informativeness term. By Assumption A.11,

$$\mathbb{E}_{a^{\mathrm{env}} \sim \pi_t}\left[\mathcal{I}^{\mathrm{sys}}(\bar{X}_t, a^{\mathrm{env}})\right] \leq \mathbb{E}_{a^{\mathrm{env}} \sim \pi_t^\star}\left[\mathcal{I}^{\mathrm{sys}}(\bar{X}_t^\star, a^{\mathrm{env}})\right] + C_{\mathrm{mis}}.$$

By Assumption A.7, we have $\mathbb{E}_{a^{\mathrm{env}} \sim \pi_t^\star}\left[\mathcal{I}^{\mathrm{sys}}(\bar{X}_t^\star, a^{\mathrm{env}})\right] \leq \Psi_{\mathrm{or}}$. Therefore,

$$\mathbb{E}_{a^{\mathrm{env}} \sim \pi_t}\left[\mathcal{I}^{\mathrm{sys}}(\bar{X}_t, a^{\mathrm{env}})\right] \leq \Psi_{\mathrm{or}} + C_{\mathrm{mis}} = K_{\mathrm{sys}}.$$

On the other hand, since $\Psi_t^{\mathrm{sys}} \geq \Psi_{\mathrm{CSR}} \geq \Psi_0^{\mathrm{sys}}$, Assumption A.10 gives

$$c^{\mathrm{sys}}_{\Theta,\mathcal{M}}(X_t) \geq m_{\Theta,\mathcal{M}}\Psi_t^{\mathrm{sys}} - c_0^{\mathrm{sys}}.$$

Substituting the two bounds into the progress identity yields

$$\begin{aligned}\mathcal{P}^{\mathrm{sys}}_{\Theta,\mathcal{M}}(X_t) &\le K_{\mathrm{sys}} - (m_{\Theta,\mathcal{M}}\Psi^{\mathrm{sys}}_t - c^{\mathrm{sys}}_0)\\ &= K_{\mathrm{sys}} + c^{\mathrm{sys}}_0 - m_{\Theta,\mathcal{M}}\Psi^{\mathrm{sys}}_t.\end{aligned}$$

Since $\Psi^{\mathrm{sys}}_t \ge \frac{K_{\mathrm{sys}}+c^{\mathrm{sys}}_0}{m_{\Theta,\mathcal{M}}}$, we obtain $\mathcal{P}^{\mathrm{sys}}_{\Theta,\mathcal{M}}(X_t) \le 0$. Finally, by Definition A.5,

$$\mathbb{E}\left[\Psi^{\mathrm{sys}}_{t+1} \mid \mathcal{F}^{\mathrm{act}}_t\right] = \Psi^{\mathrm{sys}}_t - \mathcal{P}^{\mathrm{sys}}_{\Theta,\mathcal{M}}(X_t).$$

Thus, $\mathbb{E}\left[\Psi^{\mathrm{sys}}_{t+1} \mid \mathcal{F}^{\mathrm{act}}_t\right] \ge \Psi^{\mathrm{sys}}_t$.

If the threshold region $\mathcal{R}^{\mathrm{sys}}_{\Theta,\mathcal{M}}(\Psi_{\mathrm{CSR}})$ is absorbing under the actual multi-agent update, then for every reachable state in this region the preceding argument gives non-positive expected progress, and the absorbing condition holds by assumption. Hence this region satisfies both conditions in Definition A.8, and is a shared-memory-mediated collaborative stagnation region. □

Proposition B.4 (Upper bound on the expected-potential threshold crossing index). *Suppose Assumptions A.6, A.7, A.2, and A.11 hold, and Assumption A.10 is strengthened to hold globally in the policy-averaged sense, namely $\Psi^{\mathrm{sys}}_0 = 0$. Let Ψ_{CSR} be the threshold in Proposition B.3, and define the expected-potential threshold crossing index by $\bar{t}^{\mathrm{sys}}_S := \inf\{t \ge 1 : \mathbb{E}[\Psi^{\mathrm{sys}}_t] \ge \Psi_{\mathrm{CSR}}\}$, with the convention that $\inf\emptyset = \infty$. This is a deterministic index defined by the sequence of expectations; it is not a stopping time with respect to $\mathcal{F}^{\mathrm{act}}_t$. Define*

$$\Psi^{\mathrm{sys},\star}_t := \Psi^{\mathrm{sys}}(\bar{X}^\star_t), \qquad \Xi_t := \mathbb{E}\left[\Psi^{\mathrm{sys}}_t - \Psi^{\mathrm{sys},\star}_t\right].$$

Here Ξ_t is the trajectory-level expected potential gap between the actual system and the independently evolving oracle reference. Unlike $c^{\mathrm{sys}}_{\Theta,\mathcal{M}}(X_t)$, it is not a one-step counterfactual comparison under the same realized action–observation pair. Suppose that $\Xi_1 \ge 0$. If there exists $\mu > 0$ such that, for all $t < \bar{t}^{\mathrm{sys}}_S$, $\Psi^{\mathrm{sys},\star}_t \ge \mu$ almost surely, and $\zeta_{\mathrm{sys}} := m_{\Theta,\mathcal{M}}\mu - (c^{\mathrm{sys}}_0 + C_{\mathrm{mis}}) > 0$, then

$$\bar{t}^{\mathrm{sys}}_S \le 1 + \max\left\{0,\ \left\lceil \log_{1+m_{\Theta,\mathcal{M}}} \frac{m_{\Theta,\mathcal{M}}\Psi_{\mathrm{CSR}} + \zeta_{\mathrm{sys}}}{m_{\Theta,\mathcal{M}}\Xi_1 + \zeta_{\mathrm{sys}}} \right\rceil\right\}.$$

Proof. For notational simplicity, write

$$m := m_{\Theta,\mathcal{M}}, \qquad c_0 := c^{\mathrm{sys}}_0, \qquad C := C_{\mathrm{mis}}, \qquad \zeta := \zeta_{\mathrm{sys}}.$$

Let $\pi_t := \pi^{\mathrm{env}}_{\Theta,\mathcal{M}}(\cdot \mid X_t), \pi^\star_t := \pi^{\mathrm{env}}_{\mathrm{or}}(\cdot \mid \bar{X}^\star_t)$. By Definition A.5 and the policy-averaged progress identity,

$$\mathcal{P}^{\mathrm{sys}}_{\Theta,\mathcal{M}}(X_t) = \mathbb{E}_{a^{\mathrm{env}}\sim\pi_t}\left[\mathcal{I}^{\mathrm{sys}}(\bar{X}_t, a^{\mathrm{env}})\right] - c^{\mathrm{sys}}_{\Theta,\mathcal{M}}(X_t),$$

the actual multi-agent system satisfies the following identity. Indeed, under the product coupling specified above, conditioning additionally on the independent oracle history does not change the actual round-t transition kernel given $\mathcal{F}^{\mathrm{act}}_t$:

$$\mathbb{E}\left[\Psi^{\mathrm{sys}}_{t+1} \mid \mathcal{G}_t\right] = \Psi^{\mathrm{sys}}_t - \mathbb{E}_{a^{\mathrm{env}}\sim\pi_t}\left[\mathcal{I}^{\mathrm{sys}}(\bar{X}_t, a^{\mathrm{env}})\right] + c^{\mathrm{sys}}_{\Theta,\mathcal{M}}(X_t).$$

For the oracle system, by the definition of system-level oracle informativeness,

$$\mathbb{E}\left[\Psi^{\mathrm{sys},\star}_{t+1} \mid \mathcal{G}_t\right] = \Psi^{\mathrm{sys},\star}_t - \mathbb{E}_{a^{\mathrm{env}}\sim\pi^\star_t}\left[\mathcal{I}^{\mathrm{sys}}(\bar{X}^\star_t, a^{\mathrm{env}})\right].$$

Subtracting the two identities gives

$$\begin{aligned}&\mathbb{E}\left[\Psi^{\mathrm{sys}}_{t+1} - \Psi^{\mathrm{sys},\star}_{t+1} \mid \mathcal{G}_t\right] - \left(\Psi^{\mathrm{sys}}_t - \Psi^{\mathrm{sys},\star}_t\right)\\ =\ &\mathbb{E}_{a^{\mathrm{env}}\sim\pi^\star_t}\left[\mathcal{I}^{\mathrm{sys}}(\bar{X}^\star_t, a^{\mathrm{env}})\right] - \mathbb{E}_{a^{\mathrm{env}}\sim\pi_t}\left[\mathcal{I}^{\mathrm{sys}}(\bar{X}_t, a^{\mathrm{env}})\right] + c^{\mathrm{sys}}_{\Theta,\mathcal{M}}(X_t).\end{aligned}$$

By Assumption A.11,

$$\mathbb{E}_{a^{\mathrm{env}}\sim\pi^\star_t}\left[\mathcal{I}^{\mathrm{sys}}(\bar{X}^\star_t, a^{\mathrm{env}})\right] - \mathbb{E}_{a^{\mathrm{env}}\sim\pi_t}\left[\mathcal{I}^{\mathrm{sys}}(\bar{X}_t, a^{\mathrm{env}})\right] \ge -C.$$

Since Assumption A.10 is strengthened to hold globally, we have

$$\begin{aligned}c^{\mathrm{sys}}_{\Theta,\mathcal{M}}(X_t) &\ge m\Psi^{\mathrm{sys}}_t - c_0\\ &= m\left(\Psi^{\mathrm{sys}}_t - \Psi^{\mathrm{sys},\star}_t\right) + m\Psi^{\mathrm{sys},\star}_t - c_0.\end{aligned}$$

Combining the last two inequalities yields

$$\begin{aligned}&\mathbb{E}\left[\Psi_{t+1}^{\mathrm{sys}}-\Psi_{t+1}^{\mathrm{sys},\star}\mid\mathcal{G}_t\right]\\ \geq\;&(1+m)\left(\Psi_t^{\mathrm{sys}}-\Psi_t^{\mathrm{sys},\star}\right)+m\Psi_t^{\mathrm{sys},\star}-(c_0+C).\end{aligned}$$

For every $t<\bar{t}_S^{\mathrm{sys}}$, the condition $\Psi_t^{\mathrm{sys},\star}\geq\mu$ implies

$$m\Psi_t^{\mathrm{sys},\star}-(c_0+C)\geq m\mu-(c_0+C)=\zeta.$$

Therefore, for all $t<\bar{t}_S^{\mathrm{sys}}$,

$$\mathbb{E}\left[\Psi_{t+1}^{\mathrm{sys}}-\Psi_{t+1}^{\mathrm{sys},\star}\right]\geq(1+m)\mathbb{E}\left[\Psi_t^{\mathrm{sys}}-\Psi_t^{\mathrm{sys},\star}\right]+\zeta.$$

By the definition of Ξ_t, this becomes $\Xi_{t+1}\geq(1+m)\Xi_t+\zeta$. Iterating the recursion gives, for any integer h satisfying $1\leq h<\bar{t}_S^{\mathrm{sys}}$,

$$\begin{aligned}\Xi_h&\geq(1+m)^{h-1}\Xi_1+\zeta\sum_{\ell=1}^{h-1}(1+m)^{h-1-\ell}\\ &=(1+m)^{h-1}\Xi_1+\zeta\frac{(1+m)^{h-1}-1}{m}.\end{aligned}$$

Since $\Psi_h^{\mathrm{sys}}=\Psi_h^{\mathrm{sys},\star}+\left(\Psi_h^{\mathrm{sys}}-\Psi_h^{\mathrm{sys},\star}\right)$ and $\Psi_h^{\mathrm{sys},\star}\geq0$, we have $\mathbb{E}\left[\Psi_h^{\mathrm{sys}}\right]\geq\Xi_h$. Thus, a sufficient condition for $\mathbb{E}[\Psi_h^{\mathrm{sys}}]\geq\Psi_{\mathrm{CSR}}$ is

$$(1+m)^{h-1}\Xi_1+\zeta\frac{(1+m)^{h-1}-1}{m}\geq\Psi_{\mathrm{CSR}}.$$

Rearranging gives

$$(1+m)^{h-1}\geq\frac{m\Psi_{\mathrm{CSR}}+\zeta}{m\Xi_1+\zeta}.$$

Define $h_0:=1+\max\left\{0,\left\lceil\log_{1+m}\frac{m\Psi_{\mathrm{CSR}}+\zeta}{m\Xi_1+\zeta}\right\rceil\right\}$. If $h_0<\bar{t}_S^{\mathrm{sys}}$, then the preceding lower bound applies at $h=h_0$ and gives $\mathbb{E}[\Psi_{h_0}^{\mathrm{sys}}]\geq\Psi_{\mathrm{CSR}}$, contradicting the definition of $\bar{t}_S^{\mathrm{sys}}$. Therefore, $\bar{t}_S^{\mathrm{sys}}\leq h_0$. Substituting back $m=m_{\Theta,\mathcal{M}},\quad\zeta=\zeta_{\mathrm{sys}}$, gives the desired bound. □

The following theorem combines the two preceding results. Its first part establishes persistence of the collaborative stagnation region after pathwise threshold entry, while its second part gives an expectation-level threshold-crossing bound under the stronger conditions of Proposition B.4.

Theorem B.5 (Shared-memory-mediated collaborative stagnation region and expected-potential threshold crossing). *Suppose Assumptions A.6, A.7, A.2, A.10, and A.11 hold in the policy-averaged sense. Define* $\Psi_{\mathrm{CSR}}:=\max\left\{\Psi_0^{\mathrm{sys}},\frac{\Psi_{\mathrm{or}}+C_{\mathrm{mis}}+c_0^{\mathrm{sys}}}{m_{\Theta,\mathcal{M}}}\right\}$, *and let* $\mathcal{R}_{\Theta,\mathcal{M}}^{\mathrm{sys}}(\Psi_{\mathrm{CSR}}):=\{X\in\mathcal{X}_{\mathrm{sys}}:\Psi^{\mathrm{sys}}(X)\geq\Psi_{\mathrm{CSR}}\}$. *Assume that* $\mathcal{R}_{\Theta,\mathcal{M}}^{\mathrm{sys}}(\Psi_{\mathrm{CSR}})$ *is absorbing under the actual policy-averaged multi-agent update, and define the pathwise threshold entry time* $t_S^{\mathrm{sys}}:=\inf\{t\geq1:\Psi_t^{\mathrm{sys}}\geq\Psi_{\mathrm{CSR}}\}$. *If* $t_S^{\mathrm{sys}}<\infty$, *then for every* $t\geq t_S^{\mathrm{sys}},\mathbb{E}\left[\Psi_{t+1}^{\mathrm{sys}}\mid\mathcal{F}_t^{\mathrm{act}}\right]\geq\Psi_t^{\mathrm{sys}}$. *Thus, once the actual trajectory enters the threshold region, it remains in a shared-memory-mediated collaborative stagnation region with non-positive expected system-level progress.*

Moreover, suppose that Assumption A.10 is strengthened to hold globally in the policy-averaged sense, namely $\Psi_0^{\mathrm{sys}}=0$. *Define the expected-potential threshold crossing index* $\bar{t}_S^{\mathrm{sys}}:=\inf\{t\geq1:\mathbb{E}\left[\Psi_t^{\mathrm{sys}}\right]\geq\Psi_{\mathrm{CSR}}\}$ *with the convention that* $\inf\emptyset=\infty$. *This is a deterministic index defined by the sequence of expectations, not a stopping time with respect to* $\mathcal{F}_t^{\mathrm{act}}$. *Let*

$$\Psi_t^{\mathrm{sys},\star}:=\Psi^{\mathrm{sys}}(\bar{X}_t^\star),\qquad\Xi_t:=\mathbb{E}\left[\Psi_t^{\mathrm{sys}}-\Psi_t^{\mathrm{sys},\star}\right].$$

Here Ξ_t *has the same trajectory-level expected-gap interpretation as above. Assume that* $\Xi_1\geq0$ *and that there exists* $\mu>0$ *such that, for all* $t<\bar{t}_S^{\mathrm{sys}}$, $\Psi_t^{\mathrm{sys},\star}\geq\mu$ *almost surely. If* $\zeta_{\mathrm{sys}}:=m_{\Theta,\mathcal{M}}\mu-(c_0^{\mathrm{sys}}+C_{\mathrm{mis}})>0$, *then* $\bar{t}_S^{\mathrm{sys}}\leq1+\max\left\{0,\left\lceil\log_{1+m_{\Theta,\mathcal{M}}}\frac{m_{\Theta,\mathcal{M}}\Psi_{\mathrm{CSR}}+\zeta_{\mathrm{sys}}}{m_{\Theta,\mathcal{M}}\Xi_1+\zeta_{\mathrm{sys}}}\right\rceil\right\}$. *This second bound concerns the deterministic expectation-level threshold crossing index and does not imply a pathwise upper bound on* t_S^{sys}.

Proof. By Proposition B.3, every reachable state satisfying $\Psi_t^{\mathrm{sys}} \geq \Psi_{\mathrm{CSR}}$ has non-positive policy-averaged system-level progress: $\mathcal{P}^{\mathrm{sys}}_{\Theta,\mathcal{M}}(X_t) \leq 0$. Using Definition A.5, this is equivalent to

$$\mathbb{E}\left[\Psi_{t+1}^{\mathrm{sys}} \mid \mathcal{F}_t^{\mathrm{act}}\right] = \Psi_t^{\mathrm{sys}} - \mathcal{P}^{\mathrm{sys}}_{\Theta,\mathcal{M}}(X_t) \geq \Psi_t^{\mathrm{sys}}.$$

If $t_S^{\mathrm{sys}} < \infty$, then $X_{t_S^{\mathrm{sys}}} \in \mathcal{R}^{\mathrm{sys}}_{\Theta,\mathcal{M}}(\Psi_{\mathrm{CSR}})$. Since this region is absorbing, we have $X_t \in \mathcal{R}^{\mathrm{sys}}_{\Theta,\mathcal{M}}(\Psi_{\mathrm{CSR}})$ for every $t \geq t_S^{\mathrm{sys}}$. Hence the same non-positive-progress conclusion holds for all such t.

The bound on $\bar{t}_S^{\mathrm{sys}}$ follows directly from Proposition B.4 under the additional global one-step actual–oracle gap-growth condition $\Psi_0^{\mathrm{sys}} = 0$, the initial trajectory-level gap condition $\Xi_1 \geq 0$, the oracle lower-bound condition $\Psi_t^{\mathrm{sys},\star} \geq \mu$ for all $t < \bar{t}_S^{\mathrm{sys}}$, and the positivity condition $\zeta_{\mathrm{sys}} > 0$. This proves the theorem. □

The preceding results characterize when the shared-memory system enters and remains in a stalled regime. We next connect this post-entry behavior to the training-time critic and show how the resulting value drift propagates into full-horizon advantage estimates.

C System-Level Credit Distortion and Ideal-Boundary Preservation

C.1 Post-Entry System-Level Value Drift

Lemma C.1 (**System truth-confidence drop implies system-level value drop**). *Suppose Assumption A.16 and Assumption A.17 hold. Then, for every ℓ satisfying $t_S^{\mathrm{sys}} \leq \ell \leq T-2$,*

$$\mathbb{E}\big[V_{\ell+1}^{\mathrm{sys}} - V_\ell^{\mathrm{sys}} \,\big|\, \mathcal{F}_\ell^{\mathrm{act}}\big] \leq -\rho_V^{\mathrm{sys}}, \qquad \rho_V^{\mathrm{sys}} := \kappa_V \rho_\chi^{\mathrm{sys}}. \tag{8}$$

Proof. By Assumption A.16,

$$V_{\ell+1}^{\mathrm{sys}} - V_\ell^{\mathrm{sys}} = g(\chi_{\ell+1}^{\mathrm{sys}}) - g(\chi_\ell^{\mathrm{sys}}). \tag{9}$$

By Assumption A.17, $\chi_{\ell+1}^{\mathrm{sys}} \leq \chi_\ell^{\mathrm{sys}}$ almost surely. Using the mean value theorem pathwise, when the two endpoints differ, there exists a point c_ℓ between them such that

$$g(\chi_{\ell+1}^{\mathrm{sys}}) - g(\chi_\ell^{\mathrm{sys}}) = g'(c_\ell)\left(\chi_{\ell+1}^{\mathrm{sys}} - \chi_\ell^{\mathrm{sys}}\right) \leq \kappa_V\left(\chi_{\ell+1}^{\mathrm{sys}} - \chi_\ell^{\mathrm{sys}}\right). \tag{10}$$

Here the inequality follows from $g'(c_\ell) \geq \kappa_V$ and $\chi_{\ell+1}^{\mathrm{sys}} - \chi_\ell^{\mathrm{sys}} \leq 0$. If the endpoints coincide, the inequality holds trivially. Combining Eqs. (9) and (10) yields the required pathwise value-drop bound. Taking conditional expectation with respect to $\mathcal{F}_\ell^{\mathrm{act}}$ and using Eq. (7) gives

$$\mathbb{E}\big[V_{\ell+1}^{\mathrm{sys}} - V_\ell^{\mathrm{sys}} \,\big|\, \mathcal{F}_\ell^{\mathrm{act}}\big] \leq \kappa_V \mathbb{E}\big[\chi_{\ell+1}^{\mathrm{sys}} - \chi_\ell^{\mathrm{sys}} \,\big|\, \mathcal{F}_\ell^{\mathrm{act}}\big] \leq -\kappa_V \rho_\chi^{\mathrm{sys}} = -\rho_V^{\mathrm{sys}}. \tag{11}$$

This completes the proof. □

Remark C.2. ***(Role of the post-entry value-drop lemma).*** *Theorem B.5 identifies the CSR and its entry time, whereas Lemma C.1 applies only after entry. The lemma converts post-entry system truth-confidence deterioration into the negative system-level value drift required by the credit-assignment analysis in Theorem C.3.*

C.2 System-Level Advantage Inversion

The next theorem is stated under the conditional law given a fixed system-level entry-time event E_τ with positive probability. This avoids random summation limits in the advantage decomposition.

Theorem C.3 (**Shared-memory-mediated system-level advantage inversion**). *Assume the first-part CSR conditions of Theorem B.5: the region $\mathcal{R}^{\mathrm{sys}}_{\Theta,\mathcal{M}}(\Psi_{\mathrm{CSR}})$ is absorbing under the actual policy-averaged multi-agent update, and t_S^{sys} is the pathwise threshold entry time defined there. The additional expected-potential crossing assumptions and bound in the second part of that theorem are not required here. Fix a finite horizon T and any $\tau \in \{2, \ldots, T-1\}$. Define the event*

$$\mathsf{E}_\tau := \{t_S^{\mathrm{sys}} = \tau\}. \tag{12}$$

Assume $\mathbb{P}(\mathsf{E}_\tau) > 0$ and condition on E_τ. Assume that Assumption A.15, Assumption A.16, and Assumption A.17 hold under this conditional law, and assume that the system-level critic is bounded as

$$0 \le V_\ell^{\mathrm{sys}} \le 1, \qquad \ell = 1, \dots, T. \tag{13}$$

Let $\rho_V^{\mathrm{sys}} := \kappa_V \rho_\chi^{\mathrm{sys}}$. For any $t < \tau$, define

$$\widehat{A}_{t,\mathrm{pre}}^{\mathrm{sys}} := \sum_{j=0}^{\tau-t-1} (\gamma\lambda)^j \delta_{t+j}^{\mathrm{sys}},$$
$$\mu_{\mathrm{pre}}^{\mathrm{sys}}(t,\tau) := \mathbb{E}[\widehat{A}_{t,\mathrm{pre}}^{\mathrm{sys}} \mid \mathsf{E}_\tau],$$
$$\Gamma_{\mathrm{tail}}^{\mathrm{sys}}(t,\tau) := \sum_{j=\tau-t}^{T-t-2} (\gamma\lambda)^j,$$

with the convention that an empty sum is zero. Then

$$\mathbb{E}[\widehat{A}_t^{\mathrm{sys}} \mid \mathsf{E}_\tau] \le \mu_{\mathrm{pre}}^{\mathrm{sys}}(t,\tau) - \gamma\rho_V^{\mathrm{sys}}\Gamma_{\mathrm{tail}}^{\mathrm{sys}}(t,\tau). \tag{14}$$

Consequently, if $\Gamma_{\mathrm{tail}}^{\mathrm{sys}}(t,\tau) > 0$, a sufficient condition for advantage inversion is

$$0 < \mu_{\mathrm{pre}}^{\mathrm{sys}}(t,\tau) < \gamma\rho_V^{\mathrm{sys}}\Gamma_{\mathrm{tail}}^{\mathrm{sys}}(t,\tau) \qquad \Longrightarrow \qquad \mathbb{E}[\widehat{A}_t^{\mathrm{sys}} \mid \mathsf{E}_\tau] < 0. \tag{15}$$

Proof. Write $\vartheta := \gamma\lambda$. Decompose the GAE in Eq. (3) as

$$\widehat{A}_t^{\mathrm{sys}} = \widehat{A}_{t,\mathrm{pre}}^{\mathrm{sys}} + \sum_{j=\tau-t}^{T-t-2} \vartheta^j \delta_{t+j}^{\mathrm{sys}} + \vartheta^{T-t-1}\delta_{T-1}^{\mathrm{sys}}, \tag{16}$$

where the nonterminal sum is denoted by $\mathrm{Tail}_-^{\mathrm{sys}}(t)$. By definition, $\mathbb{E}[\widehat{A}_{t,\mathrm{pre}}^{\mathrm{sys}} \mid \mathsf{E}_\tau] = \mu_{\mathrm{pre}}^{\mathrm{sys}}(t,\tau)$. For $\ell = \tau, \dots, T-2$, the outcome-based reward condition gives $r_\ell = 0$, and therefore

$$\begin{aligned}\delta_\ell^{\mathrm{sys}} &= \gamma V_{\ell+1}^{\mathrm{sys}} - V_\ell^{\mathrm{sys}} \\ &= \gamma\left(V_{\ell+1}^{\mathrm{sys}} - V_\ell^{\mathrm{sys}}\right) - (1-\gamma)V_\ell^{\mathrm{sys}}.\end{aligned} \tag{17}$$

Since $V_\ell^{\mathrm{sys}} \ge 0$ and $\gamma \in [0,1]$, we obtain

$$\delta_\ell^{\mathrm{sys}} \le \gamma\left(V_{\ell+1}^{\mathrm{sys}} - V_\ell^{\mathrm{sys}}\right). \tag{18}$$

For every $\ell \ge \tau$, the event E_τ belongs to $\mathcal{F}_\ell^{\mathrm{act}}$, since the entry time is realized by time τ. Hence Lemma C.1 remains valid under the conditional law on E_τ. Applying Lemma C.1 gives

$$\mathbb{E}[\delta_\ell^{\mathrm{sys}} \mid \mathcal{F}_\ell^{\mathrm{act}}, \mathsf{E}_\tau] \le -\gamma\rho_V^{\mathrm{sys}}, \qquad \ell = \tau, \dots, T-2. \tag{19}$$

Therefore, by iterated conditional expectation,

$$\mathbb{E}[\mathrm{Tail}_-^{\mathrm{sys}}(t) \mid \mathsf{E}_\tau] \le -\gamma\rho_V^{\mathrm{sys}}\Gamma_{\mathrm{tail}}^{\mathrm{sys}}(t,\tau). \tag{20}$$

By Assumption A.15, the terminal residual is nonpositive in conditional expectation. Since $\mathsf{E}_\tau \in \mathcal{F}_{T-1}^{\mathrm{act}}$, the tower property gives

$$\mathbb{E}[\vartheta^{T-t-1}\delta_{T-1}^{\mathrm{sys}} \mid \mathsf{E}_\tau] \le 0. \tag{21}$$

Combining the definition of $\mu_{\mathrm{pre}}^{\mathrm{sys}}(t,\tau)$ with Eqs. (20) and (21) yields Eq. (14).

The sufficient inversion condition in Eq. (15) follows by requiring the right-hand side of Eq. (14) to be negative. □

Corollary C.4 (Ideal-boundary credit preservation). *Under the CSR-entry setup, Assumptions A.15–A.17, and the bounded-critic condition of Theorem C.3, with $\mathbb{P}(\mathsf{E}_\tau) > 0$, recall the prefix-supported advantage estimator at the system-level CSR entry time τ:*

$$\widehat{A}_{t,\mathrm{pre}}^{\mathrm{sys}} := \sum_{j=0}^{\tau-t-1} (\gamma\lambda)^j \delta_{t+j}^{\mathrm{sys}}. \tag{22}$$

Then, for every $t < \tau$,

$$\mathbb{E}[\widehat{A}^{\mathrm{sys}}_{t,\mathrm{pre}} \mid \mathsf{E}_\tau] \geq \mathbb{E}[\widehat{A}^{\mathrm{sys}}_{t} \mid \mathsf{E}_\tau] + \gamma \rho_V^{\mathrm{sys}} \Gamma^{\mathrm{sys}}_{\mathrm{tail}}(t, \tau). \tag{23}$$

Thus, setting the ideal optimization boundary at the shared-memory-mediated system-level CSR entry time excludes at least the negative nonterminal tail drift quantified by $\gamma \rho_V^{\mathrm{sys}} \Gamma^{\mathrm{sys}}_{\mathrm{tail}}(t, \tau)$, where $\rho_V^{\mathrm{sys}} = \kappa_V \rho_\chi^{\mathrm{sys}}$.

Proof. By definition, $\mu^{\mathrm{sys}}_{\mathrm{pre}}(t, \tau) = \mathbb{E}[\widehat{A}^{\mathrm{sys}}_{t,\mathrm{pre}} \mid \mathsf{E}_\tau]$. Rearranging Eq. (14) gives Eq. (23). □

C.3 Role-Wise Policy-Gradient Factorization

Lemma C.5 (Role-wise system-advantage factorization). *Let $Y_t = (Y_{1,t}, \ldots, Y_{n,t})$ collect the token segments generated by the roles at round t. Suppose that the role policies have disjoint parameter blocks $\Theta = (\theta_1, \ldots, \theta_n)$ for their trainable role-specific adapters; a frozen base model may be shared. Under the fixed sequential role order, let $Y_{<i,t} := (Y_{1,t}, \ldots, Y_{i-1,t})$, with $Y_{<1,t}$ empty, and suppose that the joint conditional policy admits the chain-rule factorization*

$$\pi^{\mathrm{tok}}_\Theta(Y_t \mid X_t) = \prod_{i=1}^{n} \pi^{\mathrm{tok}}_{\theta_i}(Y_{i,t} \mid X_t, Y_{<i,t}).$$

For each role, write

$$Y_{i,t} = (y_{i,t,1}, \ldots, y_{i,t,|Y_{i,t}|}), \qquad y_{i,t,<m} := (y_{i,t,1}, \ldots, y_{i,t,m-1}),$$

and assume the autoregressive token factorization

$$\pi^{\mathrm{tok}}_{\theta_i}(Y_{i,t} \mid X_t, Y_{<i,t}) = \prod_{m=1}^{|Y_{i,t}|} \pi^{\mathrm{tok}}_{\theta_i}\big(y_{i,t,m} \mid X_t, Y_{<i,t}, y_{i,t,<m}\big).$$

Then

$$\log \pi^{\mathrm{tok}}_\Theta(Y_t \mid X_t) = \sum_{i=1}^{n} \log \pi^{\mathrm{tok}}_{\theta_i}(Y_{i,t} \mid X_t, Y_{<i,t}),$$
$$\nabla_{\theta_i} \log \pi^{\mathrm{tok}}_\Theta(Y_t \mid X_t) = \nabla_{\theta_i} \log \pi^{\mathrm{tok}}_{\theta_i}(Y_{i,t} \mid X_t, Y_{<i,t}).$$

Moreover,

$$\log \pi^{\mathrm{tok}}_{\theta_i}(Y_{i,t} \mid X_t, Y_{<i,t}) = \sum_{m=1}^{|Y_{i,t}|} \log \pi^{\mathrm{tok}}_{\theta_i}\big(y_{i,t,m} \mid X_t, Y_{<i,t}, y_{i,t,<m}\big),$$
$$\nabla_{\theta_i} \log \pi^{\mathrm{tok}}_\Theta(Y_t \mid X_t) = \sum_{m=1}^{|Y_{i,t}|} \nabla_{\theta_i} \log \pi^{\mathrm{tok}}_{\theta_i}\big(y_{i,t,m} \mid X_t, Y_{<i,t}, y_{i,t,<m}\big).$$

Consequently, a shared system-level advantage decomposes across the role-wise and token-wise policy gradients:

$$\widehat{A}^{\mathrm{sys}}_t \nabla_\Theta \log \pi^{\mathrm{tok}}_\Theta(Y_t \mid X_t) = \sum_{i=1}^{n} \sum_{m=1}^{|Y_{i,t}|} \widehat{A}^{\mathrm{sys}}_t \nabla_{\theta_i} \log \pi^{\mathrm{tok}}_{\theta_i}\big(y_{i,t,m} \mid X_t, Y_{<i,t}, y_{i,t,<m}\big),$$

where the sum is understood over the disjoint role-parameter blocks. Thus, using one system-level credit signal for all roles does not imply shared trainable actor-adapter parameters or a joint actor update. Each token score acts only on the parameter block of its generating role. No conditional-independence assumption among the role token segments given X_t is required.

Proof. The first identity follows by taking the logarithm of the sequential chain-rule factorization. When evaluating the trajectory log likelihood, the realized predecessor segments $Y_{<i,t}$ are conditioning variables. Differentiating with respect to the disjoint block θ_i therefore removes all terms whose policy kernels are parameterized by other blocks, which gives the second identity. Within

role i, the autoregressive factorization expresses the segment likelihood as a product of conditional token likelihoods. Taking its logarithm gives the token log-likelihood sum. Differentiating this sum with respect to θ_i gives the token-level score decomposition. Multiplying the resulting block-wise token gradients by the common scalar $\widehat{A}_t^{\mathrm{sys}}$ yields the final decomposition. The same argument applies when this scalar is replaced by the prefix-supported estimator $\widehat{A}_{t,<H}^{\mathrm{sys}}$ used in the role-wise PHANTOM-CONTROL objective. □

The lemma establishes role-wise and token-wise score factorization. The token-wise clipped PPO surrogate is adopted as the optimization objective; it is not claimed to be equivalent to clipping a segment-level probability ratio.

The improvement bound is specific to the ideal boundary $H = \tau$; the observable-proxy analysis below treats the practical boundary $H = \widehat{H}$ through fixed-window false-trigger control.

D OBSERVABLE-BOUNDARY ANALYSIS

Lemma D.1 (**Single-step lower-tail control for the system-level proxy**). *Let $D_t^{\mathrm{sys}} = G_t^{\mathrm{sys}} + \beta_t^{\mathrm{sys}} + \xi_t^{\mathrm{sys}}$. Assume that, on the step under consideration,*

$$G_t^{\mathrm{sys}} \geq \rho > 0 \quad a.s., \qquad |\beta_t^{\mathrm{sys}}| \leq M_d \quad a.s.,$$

and that ξ_t^{sys} is centered σ-sub-Gaussian for some $\sigma > 0$, namely

$$\mathbb{E}[\exp(\nu \xi_t^{\mathrm{sys}})] \leq \exp\left(\frac{\sigma^2 \nu^2}{2}\right), \qquad \forall \nu \in \mathbb{R}.$$

If $\Delta_{\min} < \rho - M_d$, and we set $g_{\mathrm{FT}} := \rho - M_d - \Delta_{\min} > 0$, then

$$\mathbb{P}(D_t^{\mathrm{sys}} \leq \Delta_{\min}) \leq \exp\left(-\frac{g_{\mathrm{FT}}^2}{2\sigma^2}\right).$$

In particular, the same bound holds when $\xi_t^{\mathrm{sys}} \sim \mathcal{N}(0, \sigma^2)$.

Proof. On the event $\{D_t^{\mathrm{sys}} \leq \Delta_{\min}\}$, we have $G_t^{\mathrm{sys}} + \beta_t^{\mathrm{sys}} + \xi_t^{\mathrm{sys}} \leq \Delta_{\min}$. Since $G_t^{\mathrm{sys}} \geq \rho$ and $\beta_t^{\mathrm{sys}} \geq -M_d$ almost surely, it follows that $\rho - M_d + \xi_t^{\mathrm{sys}} \leq \Delta_{\min}$. Hence

$$\xi_t^{\mathrm{sys}} \leq -(\rho - M_d - \Delta_{\min}) = -g_{\mathrm{FT}}.$$

Therefore, $\{D_t^{\mathrm{sys}} \leq \Delta_{\min}\} \subseteq \{\xi_t^{\mathrm{sys}} \leq -g_{\mathrm{FT}}\}$, and so $\mathbb{P}(D_t^{\mathrm{sys}} \leq \Delta_{\min}) \leq \mathbb{P}(\xi_t^{\mathrm{sys}} \leq -g_{\mathrm{FT}})$. For any $\nu > 0$, Chernoff's bound gives

$$\begin{aligned}
\mathbb{P}(\xi_t^{\mathrm{sys}} \leq -g_{\mathrm{FT}}) &= \mathbb{P}(\exp(-\nu \xi_t^{\mathrm{sys}}) \geq \exp(\nu g_{\mathrm{FT}})) \\
&\leq \exp(-\nu g_{\mathrm{FT}}) \mathbb{E}[\exp(-\nu \xi_t^{\mathrm{sys}})] \\
&\leq \exp\left(-\nu g_{\mathrm{FT}} + \frac{\sigma^2 \nu^2}{2}\right).
\end{aligned}$$

Optimizing over ν at $\nu = g_{\mathrm{FT}}/\sigma^2$ yields $\mathbb{P}(\xi_t^{\mathrm{sys}} \leq -g_{\mathrm{FT}}) \leq \exp\left(-\frac{g_{\mathrm{FT}}^2}{2\sigma^2}\right)$. Combining the last two displays proves the claim. The Gaussian case is a special case of the sub-Gaussian bound. □

Proposition D.2 (**False-trigger control on a fixed informative window**). *Fix $\Delta_{\min} \in \mathbb{R}$, a window length $k \in \mathbb{N}$, and a starting index t satisfying $1 \leq t \leq T - k$. Consider the fixed k-step window $\{t, t+1, \ldots, t+k-1\}$. Assume that, on this window, the following hold almost surely:*

(A1) ***Positive-margin system-level informative segment:***

$$G_{t+j}^{\mathrm{sys}} \geq \rho > 0, \qquad j = 0, \ldots, k-1.$$

(A2) ***System-level proxy decomposition with bounded bias:***

$$D_{t+j}^{\mathrm{sys}} = G_{t+j}^{\mathrm{sys}} + \beta_{t+j}^{\mathrm{sys}} + \xi_{t+j}^{\mathrm{sys}}, \qquad |\beta_{t+j}^{\mathrm{sys}}| \leq M_d, \qquad j = 0, \ldots, k-1.$$

(A3) ***Independent centered sub-Gaussian proxy noise:*** *the random variables*

$$\xi_t^{\mathrm{sys}}, \xi_{t+1}^{\mathrm{sys}}, \ldots, \xi_{t+k-1}^{\mathrm{sys}}$$

are independent, each centered, and each σ-sub-Gaussian for a common scale $\sigma > 0$:

$$\mathbb{E}\left[\exp\left(\nu \xi_{t+j}^{\mathrm{sys}}\right)\right] \le \exp\left(\frac{\sigma^2 \nu^2}{2}\right), \qquad \forall \nu \in \mathbb{R}, \quad \forall j = 0, \ldots, k-1.$$

If $\Delta_{\min} < \rho - M_d$, then the false-trigger probability on this fixed k-step system-level informative window satisfies

$$\mathbb{P}\left(\mathcal{W}_{t,k}^{\mathrm{sys}}\right) \le \exp\left(-\frac{k(\rho - M_d - \Delta_{\min})^2}{2\sigma^2}\right).$$

Consequently, a sufficient condition for ensuring $\mathbb{P}\left(\mathcal{W}_{t,k}^{\mathrm{sys}}\right) \le \varepsilon_{\mathrm{FT}}, \quad \varepsilon_{\mathrm{FT}} \in (0,1)$, is $k(\rho - M_d - \Delta_{\min})^2 \ge 2\sigma^2 \log(1/\varepsilon_{\mathrm{FT}})$.

Proof. Set $g_{\mathrm{FT}} := \rho - M_d - \Delta_{\min} > 0$. By Lemma D.1, for each $j = 0, \ldots, k-1$,

$$\mathbb{P}\left(D_{t+j}^{\mathrm{sys}} \le \Delta_{\min}\right) \le \exp\left(-\frac{g_{\mathrm{FT}}^2}{2\sigma^2}\right).$$

More strongly, by the proof of Lemma D.1, we have the event inclusion

$$\left\{D_{t+j}^{\mathrm{sys}} \le \Delta_{\min}\right\} \subseteq \left\{\xi_{t+j}^{\mathrm{sys}} \le -g_{\mathrm{FT}}\right\}.$$

Hence

$$\mathcal{W}_{t,k}^{\mathrm{sys}} = \bigcap_{j=0}^{k-1} \left\{D_{t+j}^{\mathrm{sys}} \le \Delta_{\min}\right\} \subseteq \bigcap_{j=0}^{k-1} \left\{\xi_{t+j}^{\mathrm{sys}} \le -g_{\mathrm{FT}}\right\}.$$

Taking probabilities and using independence of the noise variables yields

$$\begin{aligned}\mathbb{P}\left(\mathcal{W}_{t,k}^{\mathrm{sys}}\right) &\le \prod_{j=0}^{k-1} \mathbb{P}\left(\xi_{t+j}^{\mathrm{sys}} \le -g_{\mathrm{FT}}\right) \\ &\le \prod_{j=0}^{k-1} \exp\left(-\frac{g_{\mathrm{FT}}^2}{2\sigma^2}\right) = \exp\left(-\frac{k g_{\mathrm{FT}}^2}{2\sigma^2}\right).\end{aligned}$$

Substituting back $g_{\mathrm{FT}} = \rho - M_d - \Delta_{\min}$ proves

$$\mathbb{P}\left(\mathcal{W}_{t,k}^{\mathrm{sys}}\right) \le \exp\left(-\frac{k(\rho - M_d - \Delta_{\min})^2}{2\sigma^2}\right).$$

Finally, to guarantee $\mathbb{P}(\mathcal{W}_{t,k}^{\mathrm{sys}}) \le \varepsilon_{\mathrm{FT}}$, it suffices that

$$\exp\left(-\frac{k(\rho - M_d - \Delta_{\min})^2}{2\sigma^2}\right) \le \varepsilon_{\mathrm{FT}},$$

which is equivalent to

$$k(\rho - M_d - \Delta_{\min})^2 \ge 2\sigma^2 \log(1/\varepsilon_{\mathrm{FT}}).$$

This completes the proof. □

E TASKS, METRICS, AND TRAINING SETUP

Tasks and evaluation. We evaluate on five interactive active-reasoning tasks drawn from AR-Bench and Multi-Turn Puzzles Zhou et al. (2025); Badola et al. (2025). The tasks span hypothesis elimination, interactive questioning, preference elicitation, and recommendation. Table 2 summarizes the task-specific interaction and evaluation protocol while keeping the shared training setup fixed.

Task	Interaction	Output metric	$R_{\max}$	Evaluation split
CD	Binary circuit queries with candidate elimination	Exact Match	15	Validation
SP	Yes/No/Unknown questions for hidden-explanation recovery	Word-level F1	15	Test
GN	Number guesses with digit-position feedback	Exact Match	10	Test
PE	Pairwise queries for latent-preference estimation	Binary Similarity	10	Validation
MR	Preference elicitation followed by unseen-item recommendation	Exact Match (EM)	5	Validation

Table 2: Task interfaces and evaluation protocols. Here $R_{\max}$ is the configured maximum number of complete system interaction rounds; PE counts an estimate as correct when its cosine similarity to the target is at least 0.88.

Multi-agent architecture. All tasks use a three-role sequential multi-agent system with shared memory. The policy roles are *PreferenceEstimator*, *CriticFinalizer*, and *QueryPlanner*, executed in a fixed order within each interaction round. Each role conditions on the accumulated shared-memory history, and later roles can additionally use proposals produced earlier in the same round. The ordered role proposals are mapped by a fixed, non-trainable task-specific rule to one environment-facing interaction per round, after which the resulting environment observation is written back to shared memory for subsequent rounds. All roles share a frozen Qwen2.5-7B-Instruct backbone and use separate trainable low-rank adaptation (LoRA) adapters. During training, a centralized system-level value critic provides the common round-level GAE signal, while the policy updates remain role-specific and token-wise. The *CriticFinalizer* is a policy role and is therefore distinct from the centralized value critic.

Training and implementation. Training uses shared task-level terminal outcome rewards without additional process-reward shaping. Reported results are averaged over three runs with different random seeds. Our primary comparison uses the same multi-agent PPO architecture and training procedure for Vanilla PPO and PHANTOM-CONTROL; the latter differs only through the trajectory-level optimization intervention described in Section 5. Experiments are conducted in bfloat16 precision on two NVIDIA H100 PCIe GPUs with 80 GB memory per GPU.

Core implementation configuration. All roles use a frozen Qwen2.5-7B-Instruct backbone and a separate trainable LoRA adapter with rank $r = 16$, scaling factor 32, and dropout 0.05. The actor and critic learning rates are 3×10^{-6} and 1×10^{-4}, respectively. At an artificial boundary, policy optimization retains transitions with $t < \widehat{H}$, and the final retained TD residual bootstraps from $V^{\mathrm{sys}}(X_{\widehat{H}})$.

F IMPLEMENTATION OF PHANTOM-CONTROL

Practical detector. The system-level potential Ψ_t^{sys} and the latent CSR entry time are unavailable during rollout generation. In practice, PHANTOM-CONTROL therefore evaluates an observable system-level progress proxy after each complete shared-memory transition. Following the notation in Section 5, let

$$D_t^{\mathrm{sys}} = D_{\mathrm{obs}}\big(\mathcal{H}_t^{\mathrm{sys}}, \mathcal{H}_{t+1}^{\mathrm{sys}}\big)$$

denote the task-specific observable progress between two consecutive system states. A transition is classified as low progress when $D_t^{\mathrm{sys}} \le \Delta_{\min}$. To avoid reacting to isolated fluctuations, the detector requires this condition to hold for k consecutive system transitions. If the first such window starts at round t, the practical optimization boundary is placed at

$$\widehat{H} = t + k.$$

If no window satisfies the criterion, we set $\widehat{H} = T$ and retain the full trajectory.

Table 3: **Task-specific rollout and detector configuration used by PHANTOM-CONTROL.** The observable progress proxy is evaluated after each complete shared-memory system transition, and $R_{\max}$ denotes the configured maximum number of interaction rounds.

Task	$R_{\max}$	Observable progress	k
CD	15	Reduction in candidate-set size	6
SP	15	Cosine-distance change in the shared hypothesis	5
GN	10	Reduction in feasible candidate-set size	3
PE	10	Window-2 L_2 change in the shared preference estimate	2
MR	5	Shared-memory preference-update progress	2

Task-specific observable progress. The definition of D_t^{sys} follows the notion of observable progress available in each task, while the persistent low-progress criterion remains common across tasks. The detector is evaluated only after a complete multi-agent system transition rather than after an individual role output. For discrete progress measures, no observable refinement can be represented by $D_t^{\mathrm{sys}} = 0$; for continuous progress measures, low progress is identified relative to a task-specific threshold. The resulting task-specific instantiations are summarized in Table 3.

Common optimization boundary. Once a low-progress window is confirmed, the resulting $\widehat{H}$ is used as a single optimization boundary for the entire shared-memory trajectory. The confirming window itself is retained; only the continuation after the boundary is excluded from policy optimization. Accordingly, all roles use the common mask

$$m_t^{\mathrm{PC}} = \mathbf{1}\{t < \widehat{H}\}.$$

The boundary is therefore determined at the system-transition level rather than separately for individual role generations. This common boundary does not imply shared policy parameters: each role remains optimized through the tokens generated by its own policy.

Prefix-supported credit assignment. For retained transitions, system-level advantages are recomputed using the boundary $\widehat{H}$ as the GAE horizon. Thus, TD residuals occurring after the detected boundary do not enter the advantage estimates of retained actions. An artificial boundary is not treated as an environment terminal state: when $\widehat{H} < T$, the final retained TD residual continues to bootstrap from $V^{\mathrm{sys}}(X_{\widehat{H}})$. The value bootstrap is removed only when $X_{\widehat{H}}$ is genuinely terminal. Hence, PHANTOM-CONTROL does not redefine the environment reward or introduce an auxiliary process reward; it changes which future TD residuals contribute to the temporal credit assigned to retained collaborative actions.

Role-specific policy optimization. Although the optimization boundary and system-level advantage are shared across roles, policy updates remain role-specific. Let $Y_{i,t}$ denote the token segment generated by agent A_i at round t:

$$Y_{i,t} = (y_{i,t,1}, \ldots, y_{i,t,|Y_{i,t}|}),$$
$$y_{i,t,<m} := (y_{i,t,1}, \ldots, y_{i,t,m-1}).$$

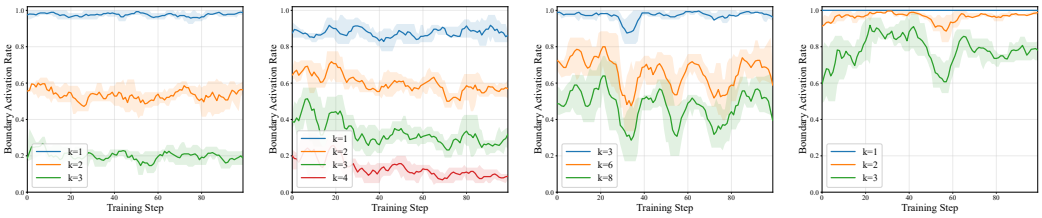
This token sequence realizes the role proposal $u_{i,t}$. With $Y_{<i,t} := (Y_{1,t}, \ldots, Y_{i-1,t})$, define the token-wise PPO ratio

$$\varrho_{i,t,m}(\theta_i) := \frac{\pi_{\theta_i}^{\mathrm{tok}}\big(y_{i,t,m} \mid X_t, Y_{<i,t}, y_{i,t,<m}\big)}{\pi_{\theta_i^{\mathrm{old}}}^{\mathrm{tok}}\big(y_{i,t,m} \mid X_t, Y_{<i,t}, y_{i,t,<m}\big)}.$$

Here, θ_i^{old} denotes the behavior-policy parameters used to generate the rollout, whereas θ_i denotes the current role-policy parameters. Conditioning on X_t denotes conditioning on the actor-observable shared-memory history represented by the system state; it does not expose the latent analytical belief variables to the acting roles.

The sequential generation order induces the factorization

$$\pi_\Theta^{\mathrm{tok}}(Y_t \mid X_t) = \prod_{i=1}^{n} \pi_{\theta_i}^{\mathrm{tok}}(Y_{i,t} \mid X_t, Y_{<i,t}), \qquad Y_t = (Y_{1,t}, \ldots, Y_{n,t}),$$



(a) CD (b) GN (c) PE (d) MR

Figure 5: **Boundary Activation Rate under different persistence settings.** The x-axis denotes PPO outer training steps, and the y-axis reports the fraction of rollouts with $\widehat{H} < T$. Different curves correspond to different persistence values k for CD, GN, PE, and MR. The solid curves show the overall training trend, while the shaded regions indicate local variation over a three-step window.

which follows directly from the conditioning order and does not assume conditional independence between role token segments.

For practical PHANTOM-CONTROL, let $H = \widehat{H}$ and $m_t^{\mathrm{PC}} = \mathbf{1}\{t < H\}$. Role i is optimized with

$$\mathcal{L}_i^{\mathrm{PC}}(\theta_i) = -\mathbb{E}\left[\sum_{t=1}^{T-1} m_t^{\mathrm{PC}} \sum_{m=1}^{|Y_{i,t}|} \min\left\{ \begin{array}{l} \varrho_{i,t,m}(\theta_i)\widehat{A}^{\mathrm{sys}}_{t,<H}, \\ \mathrm{clip}\big(\varrho_{i,t,m}(\theta_i), 1-\epsilon_{\mathrm{clip}}, 1+\epsilon_{\mathrm{clip}}\big)\widehat{A}^{\mathrm{sys}}_{t,<H} \end{array} \right\}\right],$$

where ϵ_{clip} is the PPO clipping coefficient.

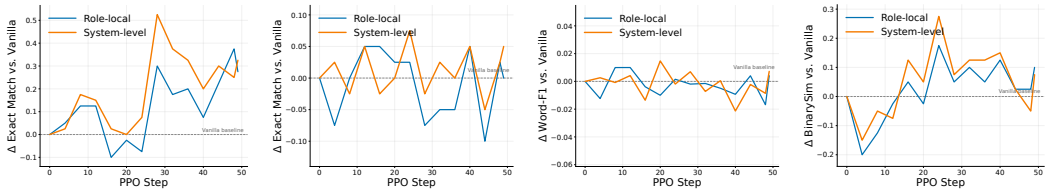
This separates the level at which the stalled continuation is controlled from the level at which policies are optimized. All roles share the common optimization boundary H, prefix mask m_t^{PC}, and system-level advantage estimate $\widehat{A}^{\mathrm{sys}}_{t,<H}$, while each token's likelihood ratio and resulting gradient remain tied to the role-specific policy that generated it. Appendix Lemma C.5 formalizes this factorization. The common system-level boundary and credit signal are compatible with role-specific policy updates.

Random-Matched control. Random-Matched is constructed independently at every PPO outer step from the boundaries produced by PHANTOM-CONTROL at that step. Let m_s be the number of rollouts assigned a boundary $H < T$ at outer step s, and let $\mathcal{L}_s = \{T - H\}$ be the corresponding multiset of masked tail lengths, measured in interaction rounds. Random-Matched samples m_s rollout indices without replacement from the full batch, randomly permutes $\mathcal{L}_s$, and assigns one length L to each selected rollout by setting $H_{\mathrm{rand}} = T - L$. Every unselected rollout keeps $H_{\mathrm{rand}} = T$. The control therefore preserves both the activated rollout count and the complete round-level masked-tail-length multiset, while removing their alignment with rollout-specific observed progress. It reassigns the realized masking lengths rather than drawing new boundaries.

After this reassignment, Random-Matched uses the same boundary-handling and optimization path as PHANTOM-CONTROL. A selected rollout uses the common mask $\mathbf{1}\{t < H_{\mathrm{rand}}\}$, prefix-supported system-level GAE, and the same artificial-boundary bootstrap from $V^{\mathrm{sys}}(X_{H_{\mathrm{rand}}})$; a genuinely terminal state still removes the bootstrap. The resulting advantages enter the same role-specific PPO objective. The paired control runs also use the same task configuration, model and LoRA setup, reward, critic, PPO hyperparameters, generation settings, training schedule, and evaluation procedure. A fixed control random seed is used only for the per-step reassignment. Thus, the controlled difference is the placement of a fixed round-level masking budget across rollouts.

G How Does Persistence Affect Boundary Activation?

Figure 5 complements the quantitative analysis in the main text by showing how the Boundary Activation Rate evolves throughout training under different persistence values k. Across CD, GN, PE, and MR, larger k consistently leads to less frequent boundary activation, and this ordering remains broadly stable despite task-specific fluctuations over PPO training. This verifies the intended operational effect of the persistence filter: requiring low observable progress over more consecutive



(a) CD (b) MR (c) SP (d) PE

Figure 6: **System-Level versus Role-Local intervention.** The x-axis denotes PPO outer training steps, and the y-axis reports the performance change relative to the Vanilla PPO baseline.

system transitions makes intervention more conservative and reduces sensitivity to short-lived fluctuations. Figure 5 therefore checks that the detector behaves as designed, while the Boundary Score and Positive-Gain Rate in Table 1 characterize the events that pass this filter. Together, they show how persistence controls which continuations are selected for system-level credit intervention.

H Does the Level of Control Matter?

Random-Matched tests whether a fixed masking budget is placed at the right points within the collaborative trajectory. The complementary Role-Local control addresses intervention scope: it contrasts control at a selected estimator update with control of the complete shared trajectory. It therefore does not isolate detector level alone.

Figure 6 compares the System-Level intervention used by PHANTOM-CONTROL with a Role-Local alternative, both relative to the same Vanilla PPO baseline. Role-Local uses the *PreferenceEstimator* as a fixed local intervention role, with the same task-specific progress criterion and detector hyperparameters as System-Level. It evaluates the estimator-local update, whereas System-Level evaluates the complete shared-memory transition and sets a common optimization boundary for the collaborative trajectory.

Role-Local confines the intervention to the selected estimator update, while rollout and system-level credit assignment remain full-horizon. System-Level instead applies a common prefix mask and recomputes prefix-supported system-level GAE, excluding post-boundary TD residuals from retained-prefix credit.

System-Level yields more persistent positive performance gains over Vanilla PPO, most clearly on CD and PE, whereas the estimator-local control is less consistent. Together with Random-Matched, this supports both targeted placement at the stalled continuation and shared-trajectory System-Level credit intervention, while policy updates remain role-specific.

Table 4: Selected notation used in the analysis and implementation.

Symbol	Meaning
Environment and system dynamics	
$S,\ s^\star,\ \Delta(S)$	Latent hypothesis space; fixed true hypothesis; PMFs over S.
$\mathcal{U}_i,\ \mathcal{A}^{\text{env}},\ \Omega$	Role-proposal, environment-facing action, and observation spaces.
$\mathcal{O},\ R,\ r_t,\ \gamma$	Observation kernel; reward function; realized reward; discount factor.
$\mathcal{E},\ \text{Agt},\ \Theta$	Active-reasoning environment; agent set; policy parameters.
$X_t = (B_t, b_t^{\text{mem}}, M_t),\ \bar{X}_t$	Full system state; belief-state projection.
$B_t,\ b_{i,t},\ b_t^{\text{mem}}$	Agent-belief collection; role belief; memory-induced belief.
$M_t,\ \mathbb{M},\ \mathcal{M}$	Memory content; memory-content space; fixed shared-memory mechanism.
$\mathcal{F}_t^{\text{act}},\ \mathcal{G}_t$	Actual pre-action filtration; joint actual–independent-oracle pre-action filtration.
$u_{i,t},\ u_{<i,t},\ U_t$	Role proposal; preceding proposals; ordered proposal sequence.

Continued on next page

Symbol	Meaning
$\pi^{\mathrm{prop}}_{\theta_i},\ \pi^{\mathrm{prop}}_{\Theta}(dU \mid X)$	Role-level proposal law; sequential joint proposal law.
$\kappa_{\mathcal{M}}(da \mid X, U),\ \pi^{\mathrm{env}}_{\Theta,\mathcal{M}}(da \mid X)$	Fixed action-selection kernel; effective environment-facing policy.
$a^{\mathrm{env}}_t,\ o_t,\ \mathfrak{B}_{\Theta}$	Environment-facing action; environment observation; actual system update.
Progress, oracle, and collaborative stagnation	
$\Psi_{i,t},\ \Psi^{\mathrm{mem}}_t,\ \Psi^{\mathrm{sys}}_t$	Agent-, memory-, and system-level log-loss potentials.
$\alpha,\ w_i,\ \omega_{\mathrm{mem}},\ \omega_i$	Memory and role aggregation weights; corresponding belief-distance weights.
$\mathcal{P}^{\mathrm{sys}}_{\Theta,\mathcal{M}}(X_t),\ G^{\mathrm{sys}}_t$	Expected and realized one-step system progress.
$\mathcal{H}^{\mathrm{sys}}_t,\ D^{\mathrm{sys}}_t$	Observable system trace; observable progress proxy.
$\mathcal{B}^{\star}_{\mathrm{Bayes}},\ \bar{\mathfrak{B}}^{\star}$	Bayesian component update; belief-only oracle update.
$\pi^{\mathrm{env}}_{\mathrm{or}},\ d_{\mathrm{bel}}$	Oracle action rule; belief-space pseudometric.
$c^{\mathrm{sys}}_{\Theta,\mathcal{M}}(X_t)$	One-step actual–oracle potential gap.
$\Psi^{\mathrm{sys},\star}_t,\ \Xi_t$	Independent-oracle potential; trajectory-level expected potential gap.
$C_{\mathrm{mis}},\ \zeta_{\mathrm{sys}}$	Oracle-informativeness mismatch bound; positive gap-growth margin.
$\mathcal{R}^{\mathrm{sys}}_{\Theta,\mathcal{M}}(\Psi_{\mathrm{CSR}})$	Collaborative Stagnation Region (CSR).
$\Psi^{\mathrm{sys}}_0,\ \Psi_{\mathrm{CSR}},\ t^{\mathrm{sys}}_S,\ \mathsf{E}_\tau$	Gap-growth threshold; CSR threshold; first hitting time; fixed-time CSR-entry event.
$\tau,\ T$	Fixed CSR-entry index under E_τ; terminal state index.
Credit assignment and optimization boundaries	
$V^{\mathrm{sys}}_t,\ \delta^{\mathrm{sys}}_t,\ \lambda$	System-level critic; TD residual; GAE trace coefficient.
$\widehat{A}^{\mathrm{sys}}_t$	Full-horizon system-level GAE.
$\widehat{A}^{\mathrm{sys}}_{t,<H}$	Prefix-supported GAE, zero-extended for $t \geq H$.
$\widehat{A}^{\mathrm{sys}}_{t,\mathrm{pre}}$	Ideal pre-entry GAE.
ρ^{sys}_V	Post-entry system-level value-drift scale.
$\mu^{\mathrm{sys}}_{\mathrm{pre}}(t,\tau)$	Conditional expected ideal pre-entry credit.
$\Gamma^{\mathrm{sys}}_{\mathrm{tail}}(t,\tau)$	Discounted nonterminal-tail mass.
$\Delta_{\min},\ k$	Progress threshold; detector-window length.
$\beta^{\mathrm{sys}}_t,\ \xi^{\mathrm{sys}}_t,\ M_d,\ \sigma$	Proxy bias and noise; bias bound and sub-Gaussian scale.
$g_{\mathrm{FT}},\ \varepsilon_{\mathrm{FT}}$	False-trigger margin; target fixed-window false-trigger probability.
$\mathcal{W}^{\mathrm{sys}}_{t,k}$	k-step low-progress event.
$H,\ \widehat{H}$	Generic and practical optimization boundaries.
m^{PC}_t	PHANTOM-CONTROL prefix mask.
Role-specific policy optimization	
$Y_{i,t},\ Y_{<i,t},\ \pi^{\mathrm{tok}}_{\theta_i},\ \pi^{\mathrm{tok}}_{\Theta}$	Role-i token segment; preceding role segments; role-level and sequential joint token laws.
$\theta^{\mathrm{old}}_i,\ \varrho_{i,t,m}(\theta_i)$	Behavior-policy parameters; token-wise PPO likelihood ratio.
ϵ_{clip}	PPO clipping coefficient.
$\mathcal{L}^{\mathrm{PC}}_i(\theta_i)$	Role-specific PHANTOM-CONTROL objective.