POMDP Interfaces in Agentic Reinforcement Learning: A Framework-Driven Review

Anonymous authors
Paper under double-blind review

Abstract

Agentic reinforcement learning extends language-model training to learning through long-horizon interaction. Existing surveys organize the field by capabilities, applications, or algorithm families, leaving less explicit how choices in memory, action abstraction, feedback, and experience collection interact. These dependencies matter because the same mechanism can improve learning in one configuration yet impair it in another. In this review, we develop a POMDP-based perspective that traces how task structure is translated into an implemented policy-gradient update. The resulting Four-Gap POMDP Framework organizes the literature around four interfaces: Belief, Granularity, Signal, and Rollout. Belief and Granularity determine the information states and decision units of the induced process; Signal and Rollout determine how its gradient is estimated. We synthesize research across these interfaces, examining their operating assumptions, characteristic failures, and six pairwise couplings through shared states, decision events, and distributions. A finite-difference formulation characterizes configuration-dependent interactions, while an exact Signal–Rollout identity separates credit error, correction residuals, and their joint contribution to update bias. Controlled ALFWorld studies show that collection lag affects task success differently under outcome-based and learned-PRM credit. This perspective connects research directions through the decision objects they jointly construct, providing a basis for comparing methods and identifying compatible designs. [1]

1 Introduction

Large language models are evolving from response-oriented generators into agents that perceive, act, and adapt through interaction. Code as Policies and CodeAct use generated code to interact with robotic and software environments (Liang et al., 2023; Wang et al., 2024b). Reinforcement learning optimizes language-model behavior through human feedback (Christiano et al., 2017; Ouyang et al., 2022) and verifiable rewards (Guo et al., 2025a). Its applications to agents include search and tool use (Jin et al., 2025), GUI action prediction (Luo et al., 2025a), and embodied navigation (Qi et al., 2025). The unit of learning thereby expands from an isolated completion to a trajectory shaped by observations, internal information, actions, and delayed feedback. Relevant history exceeds fixed context, execution traces mix several action granularities, success is often delayed, and large-scale collection increasingly decouples from policy updates.

A growing survey literature organizes this progress by agent capability and application domain (Zhang et al., 2026a), by the system roles played by foundation models (Schoepp et al., 2025), by training algorithms and infrastructure (Zhang et al., 2025a), or by feedback and temporal credit assignment (Kaufmann et al., 2025; Pignatelli et al., 2024; Zhang, 2026). These maps characterize the capabilities agents acquire and the construction of individual components. They provide less structure for analyzing how context construction, decision boundaries, feedback, and data collection interact within a policy update.

Long-context models can encode where target information occurs yet still fail to generate correct answers (Gao et al., 2024). External tool feedback changes the token distribution and gradient structure

[1] All experimental code is available in the anonymous repository: https://anonymous.4open.science/r/agentic-rl-experiments-5BF0.

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.

of multi-turn trajectories (Xue et al., 2026). Optimizing a policy against a proxy reward model can raise proxy scores and lower scores from a reference reward model (Gao et al., 2023). Stale rollouts either impair learning or remain useful under suitable control of importance weights (Noukhovitch et al., 2025; Zheng et al., 2026a). We examine these dependencies through a partially observable Markov decision process (POMDP) perspective: the history collects the information available for decisions, the action space specifies the events under policy control, rewards and values characterize long-term consequences, and the policy together with the environment induces the trajectory distribution used for optimization (Kaelbling et al., 1998). A score-function policy-gradient implementation constructs these objects computationally: it compresses interaction history into a policy context, factorizes the language–environment trajectory into decision units, converts sparse and delayed outcomes into decision-level value estimates, and maps behavior data to the current target policy. We call the mapping from a decision object of the task POMDP to the corresponding component of the implemented update an *interface*, and name the four the **Belief**, **Granularity**, **Signal**, and **Rollout** interfaces.

The four interfaces separate into two layers. Belief and Granularity specify the information available to the policy and the stochastic events it selects, inducing the decision process that the implementation optimizes. Signal constructs value estimates for that process, while Rollout specifies the behavioral origin and correction of its samples; together they determine its empirical gradient. Shared states, events, and distributions couple the four mappings, so changing one interface can alter the conditions under which the others remain valid. In the ALFWorld study, training with the same learned evaluator collapses under synchronous collection, whereas task success remains close to its initial level under lagged collection. The resulting interface axis complements capability taxonomies by tracing how system functions alter the decision process and its optimization semantics. The four gaps locate deviations between task and update; their six pairwise couplings characterize how those deviations interact.

Our main contributions are as follows.

- **An explanatory framework and reconstruction of the literature.** We reorganize policy-gradient agentic RL along the Belief, Granularity, Signal, and Rollout interfaces. Across the four interfaces, we compare the decision objects that methods modify, their operating assumptions, and characteristic failure modes. Six pairwise couplings reveal dependencies among these research directions.
- **A formal account of interface interaction.** We distinguish interfaces that induce a decision process from those that estimate its gradient. Formal results characterize control-relevant belief sufficiency, decision-unit score equivalence, and off-policy correction; an exact Signal$\times$Rollout identity isolates current-policy Signal error, residual correction error, and their interaction.
- **Mechanism-oriented diagnostics and design criteria.** Interface-specific observables and matched interventions localize failures to state construction, decision boundaries, credit, and behavior data. ALFWorld studies illustrate the role of objective anchoring in dense credit and the dependence of learned-evaluator performance on rollout conditions.

2 POMDP Interface Formalism

An agent does not directly observe the complete task state; it acts from natural-language observations, tool returns, and the interaction record, while the value of an action may be revealed only after several subsequent decisions. We use a finite-horizon episodic POMDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, T, Z, \mathcal{R}, H)$ with transition kernel T, observation kernel Z, task-reward function $\mathcal{R}$, horizon H, and history $h_t = (o_0, a_0, \ldots, o_t)$ (Kaelbling et al., 1998); the undiscounted convention matches the experiments. The framework uses the POMDP as a task-side reference and does not require estimating T or Z. In a language-agent implementation, complete histories and environment actions do not enter optimization unchanged. A history encoder ϕ maps h_{t_k} to the policy context $c_k = \phi(h_{t_k})$. A boundary rule σ defines policy-owned decision units $\bar{a}_k$ sampled from $\pi_\theta(\cdot \mid c_k)$ without intervening environment feedback. Capitalized *Belief* denotes this implemented control information state, not necessarily a Bayesian posterior. A realized unit may be a policy-owned token span or a learned latent action with its own likelihood. A semantic action groups complete strings, including termination, that are equivalent in the induced decision process, and its probability sums over these disjoint

realizations. Thus (ϕ, σ) determine the policy's conditioning information and the stochastic events under its control, inducing the policy class and decision process that the implementation actually optimizes.

For fixed (ϕ, σ), let $p_\theta^{\phi,\sigma}$ be the current-policy trajectory law. For $\tau \sim p_\theta^{\phi,\sigma}$, let $K = K(\tau)$ be its number of decision units, let $r_\ell^\star$ be the task reward accrued in interval ℓ, define $G_k = \sum_{\ell=k}^{K-1} r_\ell^\star$ and $J_{\phi,\sigma}(\theta) = \mathbb{E}_{p_\theta^{\phi,\sigma}}[G_0]$, and set $A_{\phi,\sigma}^{\pi_\theta}(c_k, \bar{a}_k) = \mathbb{E}[G_k \mid c_k, \bar{a}_k] - \mathbb{E}[G_k \mid c_k]$, which remains well defined when c_k is non-Markov. Using the score-function identity (Williams, 1992), the target gradient is

$$\nabla_\theta J_{\phi,\sigma}(\theta) = \mathbb{E}_{p_\theta^{\phi,\sigma}}\Big[\sum_{k=0}^{K-1} \nabla_\theta \log \pi_\theta(\bar{a}_k \mid c_k)\, A_{\phi,\sigma}^{\pi_\theta}(c_k, \bar{a}_k)\Big], \tag{2.1}$$

whereas the implemented empirical update can be written as

$$\hat{g} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K_i-1} \underbrace{\bar{w}_k^{(i)}}_{\textbf{Rollout}} \nabla_\theta \log \pi_\theta\big(\underbrace{\bar{a}_k^{(i)}}_{\textbf{Granularity}} \mid \underbrace{c_k^{(i)}}_{\textbf{Belief}}\big) \underbrace{\hat{A}_k^{(i)}}_{\textbf{Signal}}, \tag{2.2}$$

N is the number of sampled trajectories and $K_i = K(\tau^{(i)})$. For $k < K$, let ξ_k contain the decision unit $(c_k, \bar{a}_k)$ and every continuation, outcome, or sampled-group variable read by the k-th term. For $k \geq K$, assign ξ_k a terminal symbol † and set the score-weighted contribution to zero. Denote its unconditional behavior and target laws by q_k and p_k. Under $p_k \ll q_k$, $w_k = \mathrm{d}p_k/\mathrm{d}q_k$ is the exact ratio and $\bar{w}_k$ the implemented correction. The logged action-conditional component is $w_{\mathrm{act},k} = \pi_\theta(\bar{a}_k \mid c_k)/\mu_R(\bar{a}_k \mid c_k)$, with μ_R the recorded behavior policy. Write $s_k = \nabla_\theta \log \pi_\theta(\bar{a}_k \mid c_k)$; ν denotes a specified reference law over decision histories, and X_k the score-weighted estimand carried by term k (in Equation (2.2), $X_k = s_k \hat{A}_k$). Equations (2.1)–(2.2) isolate the score-function contribution at a fixed update state: $\hat{A}_k$ and $\bar{w}_k$ are treated as stop-gradient coefficients. If ϕ or σ shares trainable parameters with the policy, differentiable paths through state construction or boundary selection add gradient terms unless those operations are represented as stochastic policy decisions and included in the likelihood. Belief and Granularity define the policy-facing representation of the decision process; Signal constructs its score-weighted credit, and Rollout defines the behavior measure and correction used to estimate the gradient.

Table 1 identifies the discrepancy associated with each mapping in Equation (2.2) and the contract used to compare implementations. Appendix A gives the corresponding sufficiency, equivalence, support, and variance results; Section 3 develops the four literature routes.

The four interfaces in Figure 1 provide coordinates for the objects that enter the score-function estimator in Equation (2.2). Task sampling and discounting define the objective or occupancy measure; separate regularizers, parameterization, weight-space operations (Zheng et al., 2026b), and optimizer state act outside the score-weighted estimator. An auxiliary objective folded into $\hat{A}_k$ is a Signal design relative to the fixed task objective, whereas a separate auxiliary loss contributes another gradient term. Signal and Rollout are nested within the process fixed by Belief and Granularity. Appendix D gives the expanded diagnostic set.

Layer	Interface	Primitive discrepancy	Induced effect	Primary diagnostic
Process	Belief	control-relevant histories are aliased by ϕ	one-decision control loss $\mathcal{L}_\phi(\nu)$ (Appendix A.1)	context intervention and predictive sufficiency
Process	Granularity	score, credit, or correction use incompatible policy-owned events	score-weighted coefficient mismatch in Eq. (4.4)	ownership and boundary sensitivity
Update	Signal	$\epsilon_k = \hat{A}_k - A_{\phi,\sigma}^{\pi}$	$\mathbb{E}_{p_k}[s_k \epsilon_k]$	outcome agreement and score-aligned error on an identified distribution
Update	Rollout	$\bar{w}_k - w_k$	$\mathbb{E}_{q_k}[(\bar{w}_k - w_k) X_k]$	support, ratio moments, clipping mass, and correction residual

Table 1: Primitive interface discrepancies and their induced control or update effects.

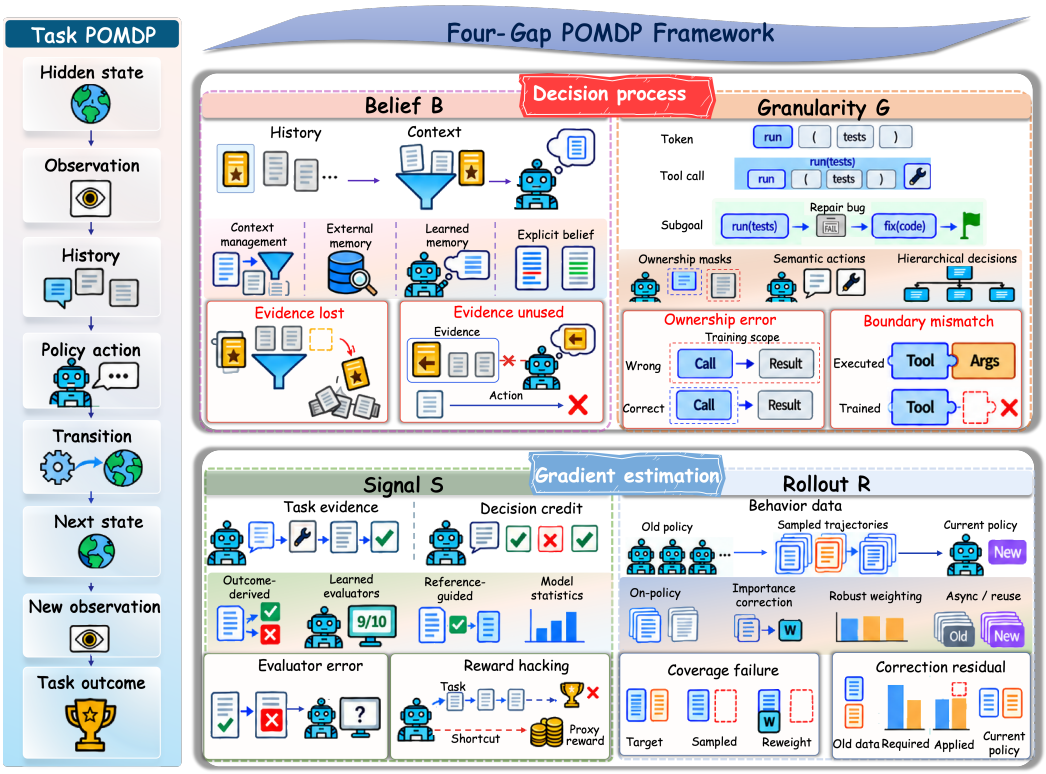

Figure 1: Agentic RL as a task POMDP and its empirical policy-gradient interface. Belief, Granularity, Signal, and Rollout map task-side reference objects to the implemented update; discrepancies in these mappings define the four interface gaps.

3 Interface Gaps in Agentic Reinforcement Learning

Reward design, memory, action abstraction, and training systems have developed as separate literatures, although each implements one of the four transformations required by Equation (2.2). Appendix C gives the extended reconstruction, and Appendix E compares method families by mechanism, characteristic failure, and coupling exposure.

3.1 Belief: Control-Relevant State Construction

In a POMDP, the posterior belief $b_t(\cdot) = \Pr(s_t \in \cdot \mid h_t)$ is sufficient for decision making under standard conditions. Agentic systems instead construct $c_k = \phi(h_{t_k})$ through history selection, summarization, retrieval, external memory, or parameter updates. A Belief gap arises when ϕ aliases histories for which no common optimal, or suitably near-optimal, action distribution exists. Approximate information states can still admit bounded control loss when they preserve the reward and transition predictions required by the task, making sufficiency relative to the task, policy, and decision boundary (Kaelbling et al., 1998; Littman et al., 2001; Subramanian et al., 2022).

Belief methods differ in where information is stored, how it is represented, and how it is updated. Managed context uses selection and summaries (Xiao et al., 2024; Wu et al., 2026). MEM1 learns to coordinate memory and reasoning (Zhou et al., 2026). Recoverable evidence may still fail to affect behavior (Liu et al., 2024; Gao et al., 2024). External memory stores evidence outside the active context and retrieves selected fragments. Errors can arise during storage, retrieval, and reintegration (Lewis et al., 2020; Packer et al., 2024; Rasmussen et al., 2025; Xu et al., 2025b). Learned memory makes those operations policy-dependent and versioned (Yan et al., 2026b; Yu et al., 2026b; Wang et al., 2024c; 2025b; Ren et al., 2026); explicit-belief representations expose hypotheses, provenance, and uncertainty to calibration tests (Lidayan et al., 2026; Singh et al., 2026; Liao et al., 2026a). Methods are compared by task-relative control loss and construction cost, with memory in-

terventions testing whether retained evidence changes action rankings as predicted (Appendices A.1 and D).

3.2 Granularity: Policy Ownership and Decision Factorization

In a POMDP, an action is a policy-selected random variable that affects the transition. Language-agent trajectories interleave policy tokens, environment responses, tool outputs, protocol markers, and multi-token operations, which need not define the same event. Granularity uses a boundary rule σ to form policy-owned units $\bar{a}_k$, fixing the event on which three operators act: the likelihood scored, the consequences credited, and the behavior–target ratio formed. A Granularity gap appears when these operators use incompatible units. For a fixed interaction record with unchanged generation-time conditioning, the score of its policy likelihood factor equals the sum of its policy-owned token scores. Sharing the same credit and correction also ensures update equivalence (Appendix A.2).

Granularity research addresses policy ownership through masking (Jiang et al., 2026a; Xue et al., 2026), defines decision units through interaction protocols (Wei et al., 2026; Liang et al., 2023; Dong et al., 2026; Luo et al., 2025a; Qi et al., 2025; Luo et al., 2025b), and develops hierarchical or learned action abstractions (Sutton et al., 1999; Zhou et al., 2024; Wen et al., 2024; Peng et al., 2026). ARIA aggregates returns over semantically clustered history–action pairs (Yang et al., 2025). GiGPO groups recurring anchor states to estimate local relative advantages (Feng et al., 2025). Both retain action boundaries and modify the Signal operator. Comparisons test policy ownership, reconstructible likelihood, credit assigned to the same event, and correction on the same joint variable. Repartitioning fixed trajectories reveals which boundaries change the gradient, the action ranking, or a resolvable consequence (Appendix D).

3.3 Signal: Objective Anchoring and Credit Construction

Fix the Belief representation ϕ and Granularity rule σ. The task reward $r^\star$ then defines the task objective $J_{\phi,\sigma}$ and current-policy advantage $A^{\pi_\theta}_{\phi,\sigma}(c_k, \bar{a}_k)$ in the induced decision process. The Signal interface applies a credit operator $\mathcal{C}_S$ to the trajectory evidence $e(\tau)$ (terminal outcomes, process verification, evaluator outputs, or policy guidance) to construct $\hat{A}_k = \mathcal{C}_S(e(\tau), c_k, \bar{a}_k)$, with error $\epsilon_k = \hat{A}_k - A^{\pi_\theta}_{\phi,\sigma}(c_k, \bar{a}_k)$. Under ideal on-policy sampling,

$$\mathbb{E}_{p_\theta^{\phi,\sigma}}\left[\sum_k s_k \hat{A}_k\right] - \nabla_\theta J_{\phi,\sigma}(\theta) = \mathbb{E}_{p_\theta^{\phi,\sigma}}\left[\sum_k s_k \epsilon_k\right]. \tag{3.1}$$

$s_k = \nabla_\theta \log \pi_\theta(\bar{a}_k \mid c_k)$. Credit need not equal the advantage pointwise: any action-independent baseline satisfies $\mathbb{E}_{\bar{a}_k \sim \pi_\theta}[s_k b(c_k)] = 0$. The Signal gap is determined by the projection of credit error onto the policy score.

Outcome verifiers and program tests anchor credit in task success (Le et al., 2022; Lambert et al., 2025; Guo et al., 2025a). Monte Carlo methods estimate continuation values from rollouts sampled at intermediate states (Kazemnejad et al., 2025; Guo et al., 2025b). Learned critics, PRMs, and judges amortize evaluation, making calibration and outcome agreement on current-policy states central (Lightman et al., 2024; Xi et al., 2026; Liu et al., 2026a; Gao et al., 2023; Zheng et al., 2025b; Song et al., 2025; Zhang et al., 2025d; Cui et al., 2026; Liu et al., 2026b; Juneja et al., 2026; Cheng et al., 2025). SPA-RL uses a learned progress estimator to attribute terminal reward to individual steps (Wang et al., 2025a). Verifiable process rewards localize external checks to intermediate decisions where correctness is testable (Yuan et al., 2026). MOPD and SDPO construct dense advantages from teacher–student log-probability differences on student-generated trajectories (Ma et al., 2026; Hübotter et al., 2026). SDAR combines task RL with a gated self-distillation loss using privileged context (Lu et al., 2026). Scalar credit can also be derived from reference information or privileged replay of the policy's own trajectory (Tao et al., 2026; Shan et al., 2026; Wang et al., 2026h; Yang et al., 2026b). Model-internal pseudo-labels and confidence statistics provide rewards without external answer labels (Zuo et al., 2025; Shafayat et al., 2025; Zhang et al., 2025b; Prabhudesai et al., 2026; Zhao et al., 2026a). Self-generated supervision can amplify errors (Wang et al., 2026c), and gains from majority-vote rewards vary across model families (Shao et al., 2026). Self-generated curricula additionally change the task distribution, with the evaluation rule remaining the Signal component (Zhao et al., 2025; Huang et al., 2026a).

Signal constructions also differ in whether they preserve the task optimum. Baselines and valid potential shaping (Ng et al., 1999) preserve the task optimum; reward penalties and scalarized objectives modify the optimization objective; separate losses contribute additional gradient terms. Across these cases, evidence provenance and calibration distribution are more informative than feedback density. Current-policy sign accuracy, score–label gap, and outcome agreement test whether denser evidence retains task meaning (Appendix D).

3.4 Rollout: Behavior Distribution

Fix ϕ, σ, and the Signal operator. For one estimator term, let ξ be the complete random element defined in Section 2. The current policy induces a target law p_θ on ξ, whereas data entering an update after collection, truncation, and selection follow a behavior law q_R on the same space. Distribution shift alone need not bias the estimator: a Rollout gap appears when target support is missing or when implemented correction leaves a residual. For an integrable estimand $X(\xi)$, the Rollout error is

$$\mathcal{E}_R(X) = \mathbb{E}_{q_R}[\bar{w}X] - \mathbb{E}_{p_\theta}[X] = \mathbb{E}_{q_R}[(\bar{w} - w)X]\,, \qquad w = \frac{\mathrm{d}p_\theta}{\mathrm{d}q_R}, \tag{3.2}$$

under the support condition, with w and $\bar{w}$ as defined in Section 2. Data generation and selection determine q_R, recorded behavior likelihood determines whether w can be reconstructed, and the correction rule determines $\bar{w}$. If selection depends on actions or outcomes, μ_R must denote the conditional law after selection; generation-time likelihood alone omits the selection contribution. With $X = s_k\hat{A}_k$, $\mathcal{E}_R(X)$ is the residual between the behavior-data update and its current-policy target. The support and second-moment conditions for exact correction appear in Appendix A.3.

Synchronous on-policy optimization refreshes rollouts frequently, limiting stale-data reuse while keeping behavior likelihood and execution state traceable over a short update cycle (Schulman et al., 2017; Shao et al., 2024). Importance correction reweights behavior-policy samples using likelihood ratios (Espeholt et al., 2018). Token-, prefix-, turn-, and sequence-level ratios act on different events and must agree with the unit carrying score and credit (Zheng et al., 2025a; Li et al., 2026a). Long autoregressive ratios create unstable tails. Weight tapering modifies sample weights, and geometric aggregation changes how weighted terms are combined (Le Roux et al., 2025; Zhao et al., 2026b). M^{2}PO uses a second-moment criterion on log importance ratios to mask token updates (Zheng et al., 2026a). GAC uses gradient projection to control persistently aligned stale gradients (Xu et al., 2026). Asynchronous systems decouple generation from learning (Noukhovitch et al., 2025; Fu et al., 2025). APRIL resumes unfinished responses after policy updates in a synchronous training loop (Zhou et al., 2025). Agentic interaction adds a provenance requirement: tool outputs, environment branches, and memory versions can alter context occupancy beyond the correction a local action ratio provides. Train–inference engine mismatch can separate the behavior policy from π_θ even at zero lag (Zhong et al., 2026). Sampling temperature, nucleus truncation, and search-guided generation can also alter the behavior distribution.

Rollout methods can be compared by their correction granularity and the residual between implemented and exact ratios on the event carrying credit. Ratio statistics and reconstruction of behavior likelihood, generating context, and system version provide the corresponding diagnostics (Appendix D).

4 Interaction Structure of the Four Interfaces

The four interfaces jointly determine the empirical policy update. Let B, G, S, R denote state construction, decision factorization, credit construction, and behavior data, with $\mathcal{G}(B, G, S, R) = \mathbb{E}_{q_R}[\bar{w}_R\, \nabla_\theta \log \pi_\theta(\bar{a}_G \mid c_B)\, \hat{A}_S(c_B, \bar{a}_G)]$. Subscripts identify the interface defining each object; all factors are evaluated under the complete configuration (B, G, S, R). Their dependence follows from shared objects: c_B conditions both policy and value, $\bar{a}_G$ carries score and credit, and q_R with $\bar{w}_R$ defines the measure of the implemented update.

4.1 Finite-Difference Structure of Interface Effects

Let $F : \{0,1\}^{\{B,G,S,R\}} \to \mathbb{R}^d$ map each interface configuration to its response, with 0 denoting the reference choice and 1 the alternative. For fixed x_{-ij}, define $D_{ij}(x_{-ij}) = F(1_i, 1_j, x_{-ij}) -$

$F(1_i, 0_j, x_{-ij}) - F(0_i, 1_j, x_{-ij}) + F(0_i, 0_j, x_{-ij})$. With $\Delta_i F(x_{-i}) = F(1_i, x_{-i}) - F(0_i, x_{-i})$ and $\Delta^U = \prod_{i \in U} \Delta_i$, the Boolean finite-difference expansion is

$$F(\mathbf{1}) - F(\mathbf{0}) = \sum_{\varnothing \neq U \subseteq \{B,G,S,R\}} (\Delta^U F)(\mathbf{0}). \tag{4.1}$$

Equation (4.1) accounts for main, pairwise, and higher-order response effects on a fixed surface; these terms are distinct from the four interface gaps. Its mechanistic interpretation requires a shared estimand; Belief or Granularity changes that alter the policy class require a bridge through task return, matched trajectories, or deterministic event regrouping (Appendix A.4). The six structural pairings separate by shared object: c_B links Signal×Belief, Belief×Granularity, and Belief×Rollout; $\bar{a}_G$ links Signal×Granularity and Granularity×Rollout; and the behavior measure links Signal×Rollout. A method has coupling exposure when its mechanism perturbs one of these shared objects; an empirical interaction is the corresponding non-additive finite difference measured in matched configurations.

4.2 Conditioning-State Couplings

The conditioning state $c_B = \phi(h)$ determines the value distinctions representable from c_B, the semantics of a decision unit, and state occupancy under behavior. For Signal×Belief, let $(h, \bar{a}) = (h_{t_k}, \bar{a}_k)$ and define $A^\pi(h, \bar{a}) = \mathbb{E}[G_k \mid h, \bar{a}] - \mathbb{E}[G_k \mid h]$. Let η be a joint law over $(h, \bar{a})$ with $A^\pi \in L^2(\eta)$, and set $c_B = \phi(h)$; the infimum ranges over $(c_B, \bar{a})$-measurable $f \in L^2(\eta)$:

$$\inf_f \mathbb{E}_\eta\left[\left(A^\pi(h, \bar{a}) - f(c_B, \bar{a})\right)^2\right] = \mathbb{E}_\eta[\mathrm{Var}(A^\pi(h, \bar{a}) \mid c_B, \bar{a})]\,. \tag{4.2}$$

The right-hand side measures value distinctions present in task history but absent from c_B. Under current-policy sampling and shared conditioning, the projection $f^\star(c_B, \bar{a}) = \mathbb{E}[A^\pi(h, \bar{a}) \mid c_B, \bar{a}]$ is the advantage of the process induced by (ϕ, σ), and its residual is orthogonal to every square-integrable $(c_B, \bar{a})$-measurable policy score. Belief compression can change the induced control problem and value-estimation variance. Fitting or applying an evaluator with different conditioning information or under a different sampling distribution can misalign its credit with the advantage induced by the policy's actual context. Methods address this interaction by contrasting matched replay views, truncating belief-trap tails, assigning credit to belief transitions, or separating belief update from action selection (Yang et al., 2026b; Zou et al., 2026b; Tang et al., 2026; Singh et al., 2026). Advantage Reweighting (AREW) adjusts stepwise advantages using directional critiques of information acquisition and belief updates (Zou et al., 2026a).

Belief×Granularity concerns how state updates relate to decision boundaries. A memory write or subgoal switch can change the context used for subsequent actions. If a coarse scored segment spans this change, its policy likelihood must include each policy-owned action under its generation-time context. HiPER and MemPO assign credit to subgoal decisions and model-generated memory segments, respectively (Peng et al., 2026; Li et al., 2026c). MiRA uses milestone progress, MemQ propagates value through memory dependencies, and Memory-R2 compares rerollouts of a session initialized with the same memory state (Wang et al., 2026d; Liao et al., 2026b; Yan et al., 2026a).

Belief also interacts with Rollout when the conditioning state is generated by a changing retrieval, memory, or belief-update process. MemQ keeps the language model frozen and guides retrieval from an expanding memory store with learned Q-values (Liao et al., 2026b). Under the support conditions for the joint context–action event, the exact density ratio for that event factorizes as

$$\frac{p_\theta(c_B, \bar{a})}{q_R(c_B, \bar{a})} = \frac{d_\theta(c_B)}{d_R(c_B)} \frac{\pi_\theta(\bar{a} \mid c_B)}{\mu_R(\bar{a} \mid c_B)}. \tag{4.3}$$

d_θ and d_R are target and behavior context occupancies under the same decision-index sampling convention; w_{act} is the second factor only. The factorization is exact when μ_R denotes the post-selection conditional action law. If it instead denotes the generation policy and selection depends on actions or outcomes, the full ratio also contains the selection contribution; outcome-dependent selection may require extending the corrected event to ξ.

4.3 Decision-Event Couplings

Partition a realized interaction record into m successive subunits without changing policy ownership or generation-time conditioning. Let $L'_{\theta,\ell}$ be the policy likelihood factor of subunit ℓ and $L_{\theta,G} =$

$\prod_{\ell=1}^{m} L'_{\theta,\ell}$. The policy scores satisfy

$$s_G = \nabla_\theta \log L_{\theta,G} = \sum_{\ell=1}^{m} s'_\ell, \qquad s'_\ell = \nabla_\theta \log L'_{\theta,\ell}.$$

Conditional on the same initial history, the exact density ratio for the joint record satisfies $w_G = \prod_{\ell=1}^{m} w'_\ell$ by the probability chain rule. Each w'_ℓ compares the target and behavior conditional densities of the complete subunit, including environment feedback, given the preceding history.

Write $\alpha_G = \bar{w}_G \hat{A}_G$ and $\alpha'_\ell = \bar{w}'_\ell \hat{A}'_\ell$. The difference between coarse- and fine-unit updates on a fixed trajectory is

$$\alpha_G s_G - \sum_{\ell=1}^{m} \alpha'_\ell s'_\ell = \sum_{\ell=1}^{m} (\alpha_G - \alpha'_\ell)\, s'_\ell. \tag{4.4}$$

Score factorization alone does not ensure update equivalence. The updates agree when the score-weighted coefficient differences sum to zero; matching each coefficient guarantees this equality. Changing decision boundaries can change the value-comparison object, baseline, and continuation distribution used to construct advantages. Segment-, action-, hierarchy-, and graph-level methods construct local credit at their chosen boundaries, often through continuation comparison or boundary-specific value estimates (Wang et al., 2026h; Guo et al., 2025b; Wang et al., 2026a; Gao et al., 2026b; Peng et al., 2026; Wang et al., 2026f).

Granularity×Rollout changes the correction component by redefining the random event whose likelihood is corrected. Sequence-, turn-, action-, and segment-level surrogate ratios regularize tails through normalization, clipping, or aggregation, but must still match the policy-owned event (Zheng et al., 2025a; Li et al., 2026a; Wang et al., 2026a; Gao et al., 2026b; Zhao et al., 2026b). Tool-interaction masks illustrate the distinction between score ownership and distribution correction: they remove environment tokens from the policy score without removing the feedback-induced shift in later policy tokens (Jiang et al., 2026a; Xue et al., 2026).

4.4 Signal–Rollout Coupling under Behavior-Distribution Shift

Fix Belief and Granularity, write s_θ for the policy score and ϵ_θ for current-parameter Signal error, and let r_S, p_θ, and q_R be evaluator, target-policy, and behavior laws on a common random element. When the Signal is decision-unit measurable, that element is $(c, \bar{a})$; otherwise it is the complete ξ. The current-policy Signal bias is $\mathcal{B}_S(p_\theta) = \mathbb{E}_{p_\theta}[s_\theta \epsilon_\theta]$. Error measured under r_S constrains this quantity only with adequate coverage and controlled score alignment (Appendix A). Behavior data instead carry $\mathbb{E}_{q_R}[\bar{w}_R s_\theta \epsilon_\theta]$: q_R selects the errors exposed to the learner and $\bar{w}_R$ determines their contribution. If the evaluator is updated from rollouts, q_R also changes ϵ_θ.

Reward-model overoptimization and process-reward benchmarks expose proxy error outside the evaluator's training distribution (Gao et al., 2023; Zheng et al., 2025b; Song et al., 2025; Zhang et al., 2025d). Evaluator-side methods adapt the evaluator on current-policy trajectories or broaden error coverage (Cui et al., 2026; Liu et al., 2026b; Juneja et al., 2026); rollout-side methods taper importance weights or mask token updates using log-ratio second moments, while GAC stabilizes stale updates through gradient projection (Le Roux et al., 2025; Zheng et al., 2026a; Xu et al., 2026). A comparison must therefore identify r_S, p_θ, q_R, and $\bar{w}_R$, and measure evaluator behavior on an identified sampling distribution together with ratio moments, clipping mass, and task outcome. Signal×Rollout admits an exact estimator identity (Section 5); Equations (4.2)–(4.4) provide projection, occupancy-ratio, and coefficient conditions for the other shared objects.

5 Signal–Rollout Interaction in Policy-Gradient Estimation

5.1 Exact Estimator-Bias Decomposition

At a fixed update state, let p and q be the current-policy and behavior laws of the random element ξ carrying one gradient term. Let $s = \nabla_\theta \log \pi_\theta(\bar{a} \mid c)$, let A be the current-policy advantage of the induced decision process, and write the implemented Signal as $\hat{A} = A + \epsilon$. With $w = \mathrm{d}p/\mathrm{d}q$ and $\bar{w}$ as defined in Section 2, the target and implemented gradients are $g = \mathbb{E}_p[sA]$ and $\tilde{g} = \mathbb{E}_q[\bar{w} s \hat{A}]$.

Theorem 5.1 (Exact Signal–Rollout Bias Decomposition)**.** *Assume $p \ll q$ and the expectations below exist. Then*

$$\tilde{g} - g = \underbrace{\mathbb{E}_p[s\epsilon]}_{\text{current-policy Signal error}} + \underbrace{\mathbb{E}_q[(\bar{w} - w)sA]}_{\text{Rollout correction residual}} + \underbrace{\mathbb{E}_q[(\bar{w} - w)s\epsilon]}_{\text{Signal}\times\text{Rollout interaction}} . \tag{5.1}$$

The first term is credit error aligned with the policy score under the target law. The second transports correction error through the true advantage, and the third transports Signal error through the same discrepancy, preserving their signed interaction. Exact correction removes both Rollout terms in expectation; ideal synchronous sampling leaves the Signal term. The identity holds at a single update state. Across updates, Rollout further changes the state sequence on which $\mathbb{E}_p[s\epsilon]$ is evaluated, and changes ϵ itself when the evaluator is fit to rollouts. The proof is given in Appendix A.5.

5.2 Objective Anchoring and Signal–Rollout Interaction

In the ALFWorld studies, outcome-derived MC credit assigns observed terminal outcomes to decision units (Figure 2; Appendix B). The learned-PRM runs obtain credit from the evaluator's predicted step scores. At 1.5B under synchronous collection, MC-uniform reaches an endpoint success rate of 0.602, compared with zero for the stronger PRM. Synchronous learned-PRM training at this scale shows rising proxy reward with zero held-out success. At 7B, MC-dense reaches a final success rate of 1.00 under both synchronous and lagged collection.

In the matched 7B comparison, PRM training ends at zero success under synchronous collection and at 0.91 under lag 8, close to the 0.89 initialization. Outcome training reaches 0.99 and 0.95, respectively. The difference between the PRM and Outcome lag effects is $+0.665$ in checkpoint-mean success, giving the Signal×Rollout interaction defined in Section 4.1.

At the last evaluator checkpoint in the 1.5B study, PRM sign accuracy on collected rollout states averages 0.05 across synchronous runs and runs with lags 1–2. The corresponding average for lags 4 and 8 is 0.56.

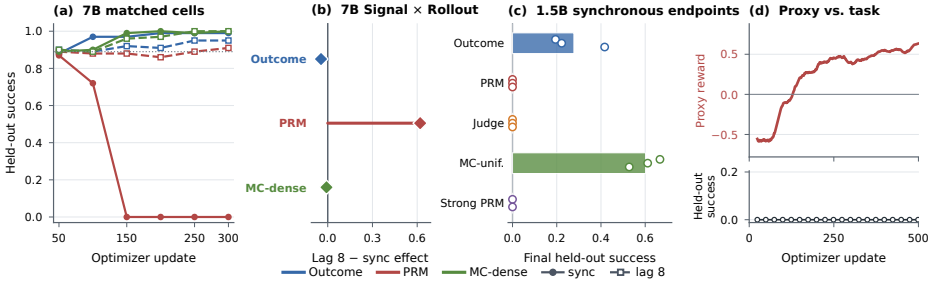

Figure 2: Task success and reward dynamics in ALFWorld. **(a)** Held-out success in six matched 7B cells; the dotted line is the common SFT baseline. **(b)** Lag-8 minus synchronous checkpoint-mean effects. **(c)** Synchronous 1.5B endpoint means and seeds: Outcome/PRM/Judge, 500 updates ($n = 3$); MC-uniform, 300 ($n = 3$); Strong PRM, 300 ($n = 2$). **(d)** PRM reward and held-out success averaged over three synchronous learned-PRM runs. The reward curve uses a 25-update moving average.

6 Conclusion

The Four-Gap POMDP Framework tracks how agentic RL implementations map a task-defined process into information states, decision events, credit estimates, and behavior correction. Shared states, decision events, and distributions connect the four research directions through six pairwise couplings. The formal analysis and ALFWorld studies show how these dependencies shape policy updates and configuration-dependent learning behavior. The interface view provides a decision-theoretic basis for comparing and designing policy-gradient agentic RL systems.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and analysis of experimental results, and to check the manuscript for errors and inconsistencies.

REFERENCES

Xin-Qiang Cai, Wei Wang, Feng Liu, Tongliang Liu, Gang Niu, and Masashi Sugiyama. Reinforcement Learning with Verifiable yet Noisy Rewards under Imperfect Verifiers. arXiv preprint, 2026. arXiv:2510.00915.

Shiyi Cao, Dacheng Li, Fangzhou Zhao, Shuo Yuan, Sumanth R. Hegde, Connor Chen, Charlie Ruan, Tyler Griggs, Shu Liu, Eric Tang, Richard Liaw, Philipp Moritz, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. SkyRL-Agent: Efficient RL Training for Multi-turn LLM Agent. arXiv preprint, 2025. arXiv:2511.16108.

Hyungjoo Chae, Sunghwan Kim, Junhee Cho, Seungone Kim, Seungjun Moon, Gyeom Hwangbo, Dongha Lim, Minjin Kim, Yeonjun Hwang, Minju Gwak, Dongwook Choi, Minseok Kang, Gwanhoon Im, ByeongUng Cho, Hyojun Kim, Jun Hee Han, Taeyoon Kwon, Minju Kim, Beongwoo Kwak, Dongjin Kang, and Jinyoung Yeo. Web-Shepherd: Advancing PRMs for Reinforcing Web Agents. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Jie Cheng, Gang Xiong, Ruixi Qiao, Lijun Li, Chao Guo, Junle Wang, Yisheng Lv, and Fei-Yue Wang. Stop Summation: Min-Form Credit Assignment Is All Process Reward Model Needs for Reasoning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, Jiarui Yuan, Huayu Chen, Kaiyan Zhang, Xingtai Lv, Shuo Wang, Yuan Yao, Xu Han, Hao Peng, Yu Cheng, Zhiyuan Liu, Maosong Sun, Bowen Zhou, and Ning Ding. Process Reinforcement through Implicit Rewards. *Transactions on Machine Learning Research*, 2026.

Guanting Dong, Hangyu Mao, Kai Ma, Licheng Bao, Yifei Chen, Zhongyuan Wang, Zhongxia Chen, Jiazhen Du, Huiyang Wang, Fuzheng Zhang, Guorui Zhou, Yutao Zhu, Ji-Rong Wen, and Zhicheng Dou. Agentic Reinforced Policy Optimization. In *International Conference on Learning Representations*, pp. 16981–17017, 2026.

Yu Du, Fangyun Wei, and Hongyang Zhang. AnyTool: Self-Reflective, Hierarchical Agents for Large-Scale API Calls. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11812–11829, 2024.

Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Vlad Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, Shane Legg, and Koray Kavukcuoglu. IMPALA: Scalable Distributed Deep-RL with Importance Weighted Actor-Learner Architectures. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1407–1416, 2018.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-Group Policy Optimization for LLM Agent Training. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Wei Fu, Jiaxuan Gao, Xujie Shen, Chen Zhu, Zhiyu Mei, Chuyi He, Shusheng Xu, Guo Wei, Jun Mei, Jiashu Wang, Tongkai Yang, Binhang Yuan, and Yi Wu. AREAL: A Large-Scale Asynchronous Reinforcement Learning System for Language Reasoning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Heyang Gao, Zexu Sun, Erxue Min, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Xu Chen. Solving the Granularity Mismatch: Hierarchical Preference Learning for Long-Horizon LLM Agents. In *International Conference on Learning Representations*, pp. 67081–67110, 2026a.

Jiaxuan Gao, Wei Fu, Minyang Xie, Shusheng Xu, Chuyi He, Zhiyu Mei, Banghua Zhu, and Yi Wu. Beyond Ten Turns: Unlocking Long-Horizon Agentic Search with Large-Scale Asynchronous RL. arXiv preprint, 2025. arXiv:2508.07976.

Lei Gao, Zhuoming Li, Mengxi Jia, Jiakang Yuan, Hongbo Sun, Hao Sun, and Xuelong Li. Segment-Aligned Policy Optimization for Multi-Modal Reasoning. In *Forty-third International Conference on Machine Learning*, 2026b.

Leo Gao, John Schulman, and Jacob Hilton. Scaling Laws for Reward Model Overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866, 2023.

Muhan Gao, TaiMing Lu, Kuai Yu, Adam Byerly, and Daniel Khashabi. Insights into LLM Long-Context Failures: When Transformers Know but Don't Tell. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7611–7625, 2024.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing. In *International Conference on Learning Representations*, pp. 57734–57811, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025a.

Yiran Guo, Lijie Xu, Jie Liu, Dan Ye, and Shuang Qiu. Segment Policy Optimization: Effective Segment-Level Credit Assignment in RL for Large Language Models. In *Advances in Neural Information Processing Systems*, volume 38, 2025b.

Chengsong Huang, Wenhao Yu, Xiaoyang Wang, Hongming Zhang, Zongxia Li, Ruosen Li, Jiaxin Huang, Haitao Mi, and Dong Yu. R-Zero: Self-Evolving Reasoning LLM from Zero Data. In *International Conference on Learning Representations*, pp. 130770–130790, 2026a.

Luke J. Huang, Zhuoyang Zhang, Qinghao Hu, Shang Yang, and Song Han. Stable Asynchrony: Variance-Controlled Off-Policy RL for LLMs. In *Forty-third International Conference on Machine Learning*, 2026b.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement Learning via Self-Distillation. In *Forty-third International Conference on Machine Learning*, 2026.

Dongfu Jiang, Yi Lu, Zhuofeng Li, Zhiheng Lyu, Ping Nie, Haozhe Wang, Alex Su, Hui Chen, Kai Zou, Chao Du, Tianyu Pang, and Wenhu Chen. VerlTool: Towards Holistic Agentic Reinforcement Learning with Tool Use. *Transactions on Machine Learning Research*, 2026a.

Haitao Jiang, Lin Ge, Hengrui Cai, and Rui Song. PABU: Progress-Aware Belief Update for Efficient LLM Agents. arXiv preprint, 2026b. arXiv:2602.09138.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan O. Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. In *Second Conference on Language Modeling*, 2025.

Gurusha Juneja, Deepak Nathani, and William Yang Wang. Adversarial Training for Process Reward Models. In *Forty-third International Conference on Machine Learning*, 2026.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998.

Minki Kang, Wei-Ning Chen, Dongge Han, Huseyin A. Inan, Lukas Wutschitz, Yanzhi Chen, Robert Sim, and Saravan Rajmohan. ACON: Optimizing Context Compression for Long-horizon LLM Agents. In *Forty-third International Conference on Machine Learning*, 2026.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. A Survey of Reinforcement Learning from Human Feedback. *Transactions on Machine Learning Research*, 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. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29557–29590, 2025.

Kimi Team, Yifan Bai, Yiping Bao, Y. Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, et al. Kimi K2: Open Agentic Intelligence. arXiv preprint, 2026. arXiv:2507.20534.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James Validad Miranda, Alisa Liu, Nouha Dziri, Xinxi Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Christopher Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing Frontiers in Open Language Model Post-Training. In *Second Conference on Language Modeling*, 2025.

Hung Le, Yue Wang, Akhilesh Deepak Gotmare, Silvio Savarese, and Steven Chu Hong Hoi. CodeRL: Mastering Code Generation through Pretrained Models and Deep Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 21314–21328, 2022.

Nicolas Le Roux, Marc G. Bellemare, Jonathan Lebensold, Arnaud Bergeron, Joshua Greaves, Alex Fréchette, Carolyne Pelletier, Eric Thibodeau-Laufer, Sándor Toth, and Sam Work. Tapered Off-Policy REINFORCE: Stable and efficient reinforcement learning for LLMs. arXiv preprint, 2025. arXiv:2503.14286.

Shiye Lei, Zhihao Cheng, and Dacheng Tao. A Step Back: Prefix Importance Ratio Stabilizes Policy Optimization. arXiv preprint, 2026. arXiv:2601.22718.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Chenliang Li, Adel Elmahdy, Alex Boyd, Zhongruo Wang, Siliang Zeng, Alfredo Garcia, Parminder Bhatia, Taha Kass-Hout, Cao Xiao, and Mingyi Hong. Stabilizing Off-Policy Training for Long-Horizon LLM Agent via Turn-Level Importance Sampling and Clipping-Triggered Normalization. arXiv preprint, 2026a. arXiv:2511.20718.

Junbo Li, Peng Zhou, Rui Meng, Meet P. Vadera, Lihong Li, and Yang Li. Turn-PPO: Turn-Level Advantage Estimation with PPO for Improved Multi-Turn RL in Agentic LLMs. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 6227–6243, 2026b.

Ruoran Li, Xinghua Zhang, Haiyang Yu, Shitong Duan, Xiang Li, Wenxin Xiang, Chonghua Liao, Xudong Guo, Yongbin Li, and Jinli Suo. MemPO: Self-Memory Policy Optimization for Long-Horizon Agents. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 23286–23301, 2026c.

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as Policies: Language Model Programs for Embodied Control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9493–9500, 2023.

Junfeng Liao, Qizhou Wang, Jianing Zhu, Bo Du, Rui Yan, and Xiuying Chen. Belief Memory: Agent Memory Under Partial Observability. arXiv preprint, 2026a. arXiv:2605.05583.

Junwei Liao, Haoting Shi, Ruiwen Zhou, Jiaqian Wang, Shengtao Zhang, Wei Zhang, Ying Wen, Zhiyu Li, Feiyu Xiong, Bo Tang, Weinan Zhang, and Muning Wen. MemQ: Integrating Q-Learning into Self-Evolving Memory Agents over Provenance DAGs. arXiv preprint, 2026b. arXiv:2605.08374.

Aly Lidayan, Jakob Bjorner, Satvik Golechha, Kartik Goyal, and Alane Suhr. ABBEL: Learning Natural-Language Belief States for Memory-Efficient Interaction. arXiv preprint, 2026. arXiv:2512.20111.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step. In *International Conference on Learning Representations*, pp. 39578–39601, 2024.

Michael L. Littman, Richard S. Sutton, and Satinder Singh. Predictive Representations of State. In *Advances in Neural Information Processing Systems*, volume 14, 2001.

Bo Liu, Chuanyang Jin, Seungone Kim, Weizhe Yuan, Wenting Zhao, Ilia Kulikov, Xian Li, Sainbayar Sukhbaatar, Jack Lanchantin, and Jason Weston. SPICE: Self-Play In Corpus Environments Improves Reasoning. arXiv preprint, 2025. arXiv:2510.24684.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Tianci Liu, Ran Xu, Tony Yu, Ilgee Hong, Carl Yang, Tuo Zhao, and Haoyu Wang. OpenRubrics: Towards Scalable Synthetic Rubric Generation for Reward Modeling and LLM Alignment. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17417–17437, 2026a.

Xiaoqian Liu, Ke Wang, Yuchuan Wu, Fei Huang, Yongbin Li, Jianbin Jiao, and Junge Zhang. Agentic Reinforcement Learning with Implicit Step Rewards. In *International Conference on Learning Representations*, pp. 129271–129291, 2026b.

Ziyan Liu, Zhezheng Hao, Yeqiu Chen, Hong Wang, Jingren Hou, Ruiyi Ding, Yongkang Yang, Wence Ji, Wei Xia, and Feng Liu. Meta-Cognitive Memory Policy Optimization for Long-Horizon LLM Agents. arXiv preprint, 2026c. arXiv:2605.30159.

Han Lu, Zichen Liu, Shaopan Xiong, Yancheng He, Wei Gao, Yanan Wu, Weixun Wang, Jiashun Liu, Yang Li, Haizhou Zhao, Ju Huang, Siran Yang, Xiaoyang Li, Yijia Luo, Zihe Liu, Ling Pan, Junchi Yan, Wei Wang, Wenbo Su, Jiamang Wang, Lin Qu, and Bo Zheng. Part II: ROLL Flash – Accelerating RLVR and Agentic Training with Asynchrony. arXiv preprint, 2025. arXiv:2510.11345.

Zhengxi Lu, Zhiyuan Yao, Zhuowen Han, Zi-Han Wang, Jinyang Wu, Qi Gu, Xunliang Cai, Weiming Lu, Jun Xiao, Yueting Zhuang, and Yongliang Shen. Self-Distilled Agentic Reinforcement Learning. arXiv preprint, 2026. arXiv:2605.15155.

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, and Abhinav Rastogi. Improve Mathematical Reasoning in Language Models by Automated Process Supervision. arXiv preprint, 2024. arXiv:2406.06592.

Run Luo, Lu Wang, Wanwei He, Longze Chen, Jiaming Li, and Xiaobo Xia. GUI-R1 : A Generalist R1-Style Vision-Language Action Model For GUI Agents. arXiv preprint, 2025a. arXiv:2504.10458.

Xufang Luo, Yuge Zhang, Zhiyuan He, Zilong Wang, Siyun Zhao, Dongsheng Li, Luna K. Qiu, and Yuqing Yang. Agent Lightning: Train ANY AI Agents with Reinforcement Learning. arXiv preprint, 2025b. arXiv:2508.03680.

Wenhan Ma, Jianyu Wei, Liang Zhao, Hailin Zhang, Bangjun Xiao, Lei Li, Qibin Yang, Bofei Gao, Yudong Wang, Rang Li, Jinhao Dong, Zhifang Sui, and Fuli Luo. MOPD: Multi-Teacher On-Policy Distillation for Capability Integration in LLM Post-Training. arXiv preprint, 2026. arXiv:2606.30406.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative Refinement with Self-Feedback. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46534–46594, 2023.

Xinji Mai, Haotian Xu, Zhong-Zhi Li, Xing W, Weinong Wang, Jian Hu, Yingying Zhang, and Wenqiang Zhang. Agent RL Scaling Law: Agent RL with Spontaneous Code Execution for Mathematical Problem Solving. arXiv preprint, 2025. arXiv:2505.07773.

Nat McAleese, Rai Michael Pokorny, Juan Felipe Ceron Uribe, Evgenia Nitishinskaya, Maja Trebacz, and Jan Leike. LLM Critics Help Catch LLM Bugs. arXiv preprint, 2024. arXiv:2407.00215.

Zijun Min, Bingshuai Liu, Ante Wang, Long Zhang, Anxiang Zeng, Haibo Zhang, and Jinsong Su. Orchestrating Tokens and Sequences: Dynamic Hybrid Policy Optimization for RLVR. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 37782–37795, 2026.

Andrew Y. Ng, Daishi Harada, and Stuart Russell. Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pp. 278–287, 1999.

Michael Noukhovitch, Shengyi Huang, Sophie Xhonneux, Arian Hosseini, Rishabh Agarwal, and Aaron Courville. Asynchronous RLHF: Faster and More Efficient Off-Policy RL for Language Models. In *International Conference on Learning Representations*, pp. 4003–4029, 2025.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as Operating Systems. arXiv preprint, 2024. arXiv:2310.08560.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22, 2023.

Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. Gorilla: Large Language Model Connected with Massive APIs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 126544–126565, 2024.

Jiangweizhi Peng, Yuanxin Liu, Ruida Zhou, Charles Fleming, Zhaoran Wang, Alfredo Garcia, and Mingyi Hong. HiPER: Hierarchical Reinforcement Learning with Explicit Credit Assignment for Large Language Model Agents. arXiv preprint, 2026. arXiv:2602.16165.

Eduardo Pignatelli, Johan Ferret, Matthieu Geist, Thomas Mesnard, Hado van Hasselt, and Laura Toni. A Survey of Temporal Credit Assignment in Deep Reinforcement Learning. *Transactions on Machine Learning Research*, 2024.

Mihir Prabhudesai, Lili Chen, Alex Ippoliti, Katerina Fragkiadaki, Hao Liu, and Deepak Pathak. Maximizing Confidence Alone Improves Reasoning. *Transactions on Machine Learning Research*, 2026.

Archiki Prasad, Alexander Koller, Mareike Hartmann, Peter Clark, Ashish Sabharwal, Mohit Bansal, and Tushar Khot. ADaPT: As-Needed Decomposition and Planning with Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 4226–4252, 2024.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. Measuring and Narrowing the Compositionality Gap in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711, 2023.

Zhangyang Qi, Zhixiong Zhang, Yizhou Yu, Jiaqi Wang, and Hengshuang Zhao. VLN-R1: Vision-Language Navigation via Reinforcement Fine-Tuning. arXiv preprint, 2025. arXiv:2506.17221.

Zekai Qu, Yinxu Pan, Ao Sun, Chaojun Xiao, and Xu Han. CoPRIS: Efficient and Stable Reinforcement Learning via Concurrency-Controlled Partial Rollout with Importance Sampling. arXiv preprint, 2025. arXiv:2511.05589.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: A Temporal Knowledge Graph Architecture for Agent Memory. arXiv preprint, 2025. arXiv:2501.13956.

Tao Ren, Weiyao Luo, Hui Yang, Rongzhi Zhu, Xiang Huang, Yuchuan Wu, Bingxue Chou, Jieping Ye, Jiafeng Liang, Yongbin Li, and Yijie Peng. Scaling Self-Evolving Agents via Parametric Memory. arXiv preprint, 2026. arXiv:2606.04536.

Sheila Schoepp, Masoud Jafaripour, Yingyue Cao, Tianpei Yang, Fatemeh Abdollahi, Shadan Golestan, Zahin Sufiyan, Osmar R. Zaiane, and Matthew E. Taylor. The Evolving Landscape of LLM- and VLM-Integrated Reinforcement Learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)*, pp. 10641–10649, 2025. Survey Track.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. arXiv preprint, 2017. arXiv:1707.06347.

Sheikh Shafayat, Fahim Tajwar, Ruslan Salakhutdinov, Jeff Schneider, and Andrea Zanette. Can Large Reasoning Models Self-Train? arXiv preprint, 2025. arXiv:2505.21444.

Yingyu Shan, Yuhang Guo, Zihao Cheng, Zeming Liu, Xiangrong Zhu, Xinyi Wang, Jiashu Yao, Wei Lin, Hongru Wang, and Heyan Huang. Learning from Own Solutions: Self-Conditioned Credit Assignment for Reinforcement Learning with Verifiable Rewards. arXiv preprint, 2026. arXiv:2606.18810.

Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, Nathan Lambert, Sewon Min, Ranjay Krishna, Yulia Tsvetkov, Hannaneh Hajishirzi, Pang Wei Koh, and Luke Zettlemoyer. Spurious Rewards: Rethinking Training Signals in RLVR. In *Forty-third International Conference on Machine Learning*, 2026.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv preprint, 2024. arXiv:2402.03300.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A Flexible and Efficient RLHF Framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 1279–1297, 2025.

Yaorui Shi, Yuxin Chen, Siyuan Wang, Sihang Li, Hengxing Cai, Qi Gu, Xiang Wang, and An Zhang. Look Back to Reason Forward: Revisitable Memory for Long-Context LLM Agents. In *International Conference on Learning Representations*, pp. 78055–78082, 2026.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8634–8652, 2023.

Joykirat Singh, Zaid Khan, Archiki Prasad, Justin Chih-Yao Chen, Akshay Nambi, Hyunji Lee, Elias Stengel-Eskin, and Mohit Bansal. Agent-BRACE: Decoupling Beliefs from Actions in Long-Horizon Tasks via Verbalized State Uncertainty. arXiv preprint, 2026. arXiv:2605.11436.

Mingyang Song, Zhaochen Su, Xiaoye Qu, Jiawei Zhou, and Yu Cheng. PRMBench: A Fine-grained and Challenging Benchmark for Process-Level Reward Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25299–25346, 2025.

Jayakumar Subramanian, Amit Sinha, Raihan Seraj, and Aditya Mahajan. Approximate Information State for Approximate Planning and Reinforcement Learning in Partially Observed Systems. *Journal of Machine Learning Research*, 23(12):1–83, 2022.

Richard S. Sutton, Doina Precup, and Satinder Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1–2):181–211, 1999.

Wenjie Tang, Minne Li, Sijie Huang, Liquan Xiao, and Yuan Zhou. Rewarding Beliefs, Not Actions: Consistency-Guided Credit Assignment for Long-Horizon Agents. arXiv preprint, 2026. arXiv:2605.20061.

Leitian Tao, Baolin Peng, Wenlin Yao, Tao Ge, Hao Cheng, Mike Hang Wang, Jianfeng Gao, and Sharon Li. TRACE: Turn-level Reward Assignment via Credit Estimation for Long-Horizon Agents. arXiv preprint, 2026. arXiv:2607.13988.

Daoyu Wang, Qingchuan Li, Mingyue Cheng, Jie Ouyang, Shuo Yu, Chunli Liu, Shijin Wang, Qi Liu, and Enhong Chen. CAPO: Critic-Guided Action-Aligned Policy Optimization for Advancing LLM Agent Capabilities. arXiv preprint, 2026a. arXiv:2604.18401.

Guoqing Wang, Sunhao Dai, Guangze Ye, Zeyu Gan, Wei Yao, Yong Deng, Xiaofeng Wu, and Zhenzhe Ying. Information Gain-based Policy Optimization: A Simple and Effective Approach for Multi-Turn Search Agents. In *International Conference on Learning Representations*, pp. 104502–104529, 2026b.

Hanlin Wang, Chak Tou Leong, Jiashuo Wang, Jian Wang, and Wenjie Li. SPA-RL: Reinforcing LLM Agents via Stepwise Progress Attribution. arXiv preprint, 2025a. arXiv:2505.20732.

Kaiqi Wang, Wenjin Hou, Yuchen Yan, Hongrui Jia, Zhisheng Zhong, Botao Ren, Yifei Chen, Songyang Zhang, Yongliang Shen, Jun Xiao, Yi Yang, Yueting Zhuang, and Hehe Fan. Diving into Reliable Self-Evolving Agents: A Survey. Preprints.org, 2026c. URL https://www.preprints.org/manuscript/202609.0913/v1. Version 1.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-Shepherd: Verify and Reinforce LLMs Step-by-step without Human Annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, 2024a.

Taiyi Wang, Sian Gooding, Florian Hartmann, Oriana Riva, and Edward Grefenstette. A Subgoal-driven Framework for Improving Long-Horizon LLM Agents. arXiv preprint, 2026d. arXiv:2603.19685.

Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. Executable Code Actions Elicit Better LLM Agents. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 50208–50232, 2024b.

Xuekang Wang, Zhuoyuan Hao, Shuo Hou, Hao Peng, Juanzi Li, and Xiaozhi Wang. Reproducing, Analyzing, and Detecting Reward Hacking in Rubric-Based Reinforcement Learning. arXiv preprint, 2026e. arXiv:2606.04923.

Yu Wang, Yifan Gao, Xiusi Chen, Haoming Jiang, Shiyang Li, Jingfeng Yang, Qingyu Yin, Zheng Li, Xian Li, Bing Yin, Jingbo Shang, and Julian McAuley. MEMORYLLM: Towards Self-Updatable Large Language Models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 50453–50466, 2024c.

Yu Wang, Dmitry Krotov, Yuanzhe Hu, Yifan Gao, Wangchunshu Zhou, Julian McAuley, Dan Gutfreund, Rogerio Feris, and Zexue He. M+: Extending MemoryLLM with Scalable Long-Term Memory. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 63308–63323, 2025b.

Yu Wang, Ryuichi Takanobu, Zhiqi Liang, Yuzhen Mao, Yuanzhe Hu, Julian McAuley, and Xiaojian Wu. Mem-α: Learning Memory Construction via Reinforcement Learning. arXiv preprint, 2025c. arXiv:2509.25911.

Yunan Wang, Minghui Song, Zihan Zhang, Shaohan Huang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. Group-Graph Policy Optimization for Long-Horizon Agentic Reinforcement Learning. arXiv preprint, 2026f. arXiv:2606.22995.

Zhenting Wang, Huancheng Chen, Jiayun Wang, and Wei Wei. Memex(RL): Scaling Long-Horizon LLM Agents via Indexed Experience Memory. arXiv preprint, 2026g. arXiv:2603.04257.

Zi-Han Wang, Zhengxi Lu, Zhiyuan Yao, Jinyang Wu, Jie Wu, Zhengzhou Cai, Yueqing Sun, Ziang Ye, Linji Hao, Qi Gu, Xunliang Cai, Yongliang Shen, and Yujiu Yang. AgentOPSD: Recursive Self-Distillation for Agentic Reinforcement Learning. arXiv preprint, 2026h. arXiv:2608.05987.

Quan Wei, Siliang Zeng, Chenliang Li, Zhongruo Wang, William Brown, Oana Frunza, Wei Deng, Anderson Schneider, Yuriy Nevmyvaka, Yang Katie Zhao, Alfredo Garcia, and Mingyi Hong. Reinforcing Multi-Turn Reasoning in LLM Agents via Fine-Grained Reward Structure and Credit Assignment. arXiv preprint, 2026. arXiv:2505.11821.

Muning Wen, Ziyu Wan, Jun Wang, Weinan Zhang, and Ying Wen. Reinforcing LLM Agents via Policy Optimization with Action Decomposition. In *Advances in Neural Information Processing Systems*, volume 37, pp. 103774–103805, 2024.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256, 1992.

Bo Wu, Sid Wang, Yunhao Tang, Jia Ding, Eryk Helenowski, Liang Tan, Tengyu Xu, Tushar Gowda, Zhengxing Chen, Chen Zhu, Xiaocheng Tang, Yundi Qian, Beibei Zhu, and Rui Hou. LlamaRL: A Distributed Asynchronous Reinforcement Learning Framework for Efficient Large-scale LLM Training. arXiv preprint, 2025a. arXiv:2505.24034.

Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason E. Weston, and Sainbayar Sukhbaatar. Meta-Rewarding Language Models: Self-Improving Alignment with LLM-as-a-Meta-Judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 11537–11554, 2025b.

Xixi Wu, Kuan Li, Yida Zhao, Liwen Zhang, Litu Ou, Huifeng Yin, Zhongwang Zhang, Xinmiao Yu, Dingchu Zhang, Yong Jiang, Pengjun Xie, Fei Huang, Minhao Cheng, Shuai Wang, Hong Cheng, and Jingren Zhou. ReSum: Unlocking Long-Horizon Search Intelligence via Context Summarization. arXiv preprint, 2026. arXiv:2509.13313.

Zhiheng Xi, Jixuan Huang, Chenyang Liao, Baodai Huang, Honglin Guo, Jiaqi Liu, Rui Zheng, Junjie Ye, Jiazheng Zhang, Wenxiang Chen, Wei He, Yiwen Ding, Guanyu Li, Zehui Chen, Zhengyin Du, Xuesong Yao, Yufei Xu, Jiecao Chen, Tao Gui, Zuxuan Wu, Qi Zhang, Xuanjing Huang, and Yu-Gang Jiang. AgentGym-RL: Training LLM Agents for Long-Horizon Decision Making through Multi-Turn Reinforcement Learning. arXiv preprint, 2025. arXiv:2509.08755.

Zhiheng Xi, Chenyang Liao, Guanyu Li, Zhihao Zhang, Wenxiang Chen, Binghai Wang, Senjie Jin, Yuhao Zhou, Jian Guan, Wei Wu, Tao Ji, Tao Gui, Qi Zhang, and Xuanjing Huang. AgentPRM: Process Reward Models for LLM Agents via Step-Wise Promise and Progress. In *Proceedings of the ACM Web Conference 2026*, pp. 4184–4195, 2026.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient Streaming Language Models with Attention Sinks. In *International Conference on Learning Representations*, pp. 21875–21895, 2024.

Yutao Xie, Nathaniel Thomas, Nicklas Hansen, Yang Fu, Li Erran Li, and Xiaolong Wang. TIPS: Turn-level Information-Potential Reward Shaping for Search-Augmented LLMs. In *International Conference on Learning Representations*, pp. 156549–156584, 2026.

Haofeng Xu, Junwei Su, Yukun Tian, Lansong Diao, Zhengping Qian, and Chuan Wu. GAC: Stabilizing Asynchronous RL Training for LLMs via Gradient Alignment Control. arXiv preprint, 2026. arXiv:2603.01501.

Peiran Xu, Zhuohao Li, Xiaoying Xing, Guannan Zhang, Debiao Li, and Kunyu Shi. Hybrid Reward Normalization for Process-supervised Non-verifiable Agentic Tasks. arXiv preprint, 2025a. arXiv:2509.25598.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-Mem: Agentic Memory for LLM Agents. In *Advances in Neural Information Processing Systems*, volume 38, 2025b.

Zhenghai Xue, Longtao Zheng, Qian Liu, Yingru Li, Xiaosen Zheng, Zejun Ma, and Bo An. SimpleTIR: End-to-End Reinforcement Learning for Multi-Turn Tool-Integrated Reasoning. In *International Conference on Learning Representations*, pp. 8424–8449, 2026.

Sikuan Yan, Ahmed Bahloul, Ercong Nie, Susanna Schwarzmann, Riccardo Trivisonno, Volker Tresp, and Yunpu Ma. Memory-R2: Fair Credit Assignment for Long-Horizon Memory-Augmented LLM Agents. arXiv preprint, 2026a. arXiv:2605.21768.

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Jinhe Bi, Kristian Kersting, Jeff Z. Pan, Hinrich Schuetze, Volker Tresp, and Yunpu Ma. Memory-R1: Enhancing Large Language Model Agents to Manage and Utilize Memories via Reinforcement Learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12805–12825, 2026b.

John Yang, Carlos Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. SWE-agent: Agent-Computer Interfaces Enable Automated Software Engineering. In *Advances in Neural Information Processing Systems*, volume 37, pp. 50528–50652, 2024.

Kun Yang, Zikang Chen, Yanmeng Wang, Zhigen Li, Ning Cheng, Shaojun Wang, and Jing Xiao. SSPO: Subsentence-level Policy Optimization. arXiv preprint, 2026a. arXiv:2511.04256.

Ruihan Yang, Yikai Zhang, Aili Chen, Xintao Wang, Siyu Yuan, Jiangjie Chen, Deqing Yang, and Yanghua Xiao. ARIA: Training Language Agents with Intention-driven Reward Aggregation. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Yi Yang, Cong Qin, Xiaodan Liu, Chishui Chen, Qing Dong, Yan Zhang, Cao Liu, Zhao Yang, Lu Pan, Jiaye Lin, and Yi Feng. Agentic Reinforcement Learning with Observation-Calibrated Self-Distillation. arXiv preprint, 2026b. arXiv:2608.04788.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations*, 2023.

Hongli Yu, Tinghong Chen, Jiangtao Feng, Jiangjie Chen, Weinan Dai, Qiying Yu, Ya-Qin Zhang, Wei-Ying Ma, Jingjing Liu, Mingxuan Wang, and Hao Zhou. MemAgent: Reshaping Long-Context LLM with Multi-Conv RL-based Memory Agent. In *International Conference on Learning Representations*, pp. 39458–39486, 2026a.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: An Open-Source LLM Reinforcement Learning System at Scale. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Yi Yu, Liuyi Yao, Yuexiang Xie, Qingquan Tan, Jiaqi Feng, Yaliang Li, and Libing Wu. Agentic Memory: Learning Unified Long-Term and Short-Term Memory Management for Large Language Model Agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21457–21483, 2026b.

Huining Yuan, Zelai Xu, Huaijie Wang, Xiangmin Yi, Jiaxuan Gao, Xiao-Ping Zhang, Yu Wang, Chao Yu, and Yi Wu. Verifiable Process Rewards for Agentic Reasoning. arXiv preprint, 2026. arXiv:2605.10325.

Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. Free Process Rewards without Process Labels. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 73511–73525, 2025.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E. Weston. Self-Rewarding Language Models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57905–57923, 2024.

Yu Yue, Yufeng Yuan, Qiying Yu, Xiaochen Zuo, Ruofei Zhu, Wenyuan Xu, Jiaze Chen, Chengyi Wang, TianTian Fan, Zhengyin Du, Xiangpeng Wei, Xiangyu Yu, Gaohong Liu, Juncai Liu, Lingjun Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Ru Zhang, Xin Liu, Mingxuan Wang, Yonghui Wu, and Lin Yan. VAPO: Efficient and Reliable Reinforcement Learning for Advanced Reasoning Tasks. arXiv preprint, 2025a. arXiv:2504.05118.

Zhongqi Yue, Weishi Wang, Yundaichuan Zhan, Juncheng Li, Daniel Dahlmeier, and Fredrik D. Johansson. Expanding the Action Space of LLMs to Reason Beyond Language. arXiv preprint, 2025b. arXiv:2510.07581.

Chenchen Zhang. From Reasoning to Agentic: Credit Assignment in Reinforcement Learning for Large Language Models. arXiv preprint, 2026. arXiv:2604.09459.

Guibin Zhang, Hejia Geng, Xiaohang Yu, Zhenfei Yin, Zaibin Zhang, Zelin Tan, Heng Zhou, Zhong-Zhi Li, Xiangyuan Xue, Yijiang Li, Yifan Zhou, Yang Chen, Chen Zhang, Yutao Fan, Zihu Wang, Songtao Huang, Francisco Piedrahita Velez, Yue Liao, Hongru Wang, Mengyue Yang, Heng Ji, Jun Wang, Shuicheng Yan, Philip Torr, and Lei Bai. The Landscape of Agentic Reinforcement Learning for LLMs: A Survey. *Transactions on Machine Learning Research*, 2026a.

Hao Zhang, Mingjie Liu, Shaokun Zhang, Songyang Han, Jian Hu, Zhenghui Jin, Yuchi Zhang, Shizhe Diao, Ximing Lu, Binfeng Xu, Zhiding Yu, Jan Kautz, and Yi Dong. ProRL Agent: Rollout-as-a-Service for RL Training of Multi-Turn LLM Agents. arXiv preprint, 2026b. arXiv:2603.18815.

Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, Yu Fu, Xingtai Lv, Yuchen Zhang, Sihang Zeng, Shang Qu, Haozhan Li, Shijie Wang, Yuru Wang, Xinwei Long, Fangfu Liu, Xiang Xu, Jiaze Ma, Xuekai Zhu, Ermo Hua, Yihao Liu, Zonglin Li, Huayu Chen, Xiaoye Qu, Yafu Li, Weize Chen, Zhenzhao Yuan, Junqi Gao, Dong Li, Zhiyuan Ma, Ganqu Cui, Zhiyuan Liu, Biqing Qi, Ning Ding, and Bowen Zhou. A Survey of Reinforcement Learning for Large Reasoning Models. arXiv preprint, 2025a. arXiv:2509.08827.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Peilin Zhao, and Yatao Bian. Right Question is Already Half the Answer: Fully Unsupervised LLM Reasoning Incentivization. In *Advances in Neural Information Processing Systems*, volume 38, 2025b.

Wenlin Zhang, Xiangyang Li, Kuicai Dong, Yichao Wang, Pengyue Jia, Xiaopeng Li, Yingyi Zhang, Derong Xu, Zhaocheng Du, Huifeng Guo, Ruiming Tang, and Xiangyu Zhao. Process vs. Outcome Reward: Which is Better for Agentic RAG Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025c.

Yaocheng Zhang, Haohuan Huang, Zijun Song, Yuanheng Zhu, Qichao Zhang, Zijie Zhao, and Dongbin Zhao. CriticSearch: Fine-Grained Credit Assignment for Search Agents via a Retrospective Critic. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 12272–12290, 2026c.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. The Lessons of Developing Process Reward Models in Mathematical Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 10495–10516, 2025d.

Zijing Zhang, Ziyang Chen, Mingxiao Li, Zhaopeng Tu, and Xiaolong Li. RLVMR: Reinforcement Learning with Verifiable Meta-Reasoning Rewards for Robust Long-Horizon Agents. In *International Conference on Learning Representations*, pp. 114676–114700, 2026d.

Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. Absolute Zero: Reinforced Self-play Reasoning with Zero Data. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. Learning to Reason without External Rewards. In *International Conference on Learning Representations*, pp. 2548–2581, 2026a.

Yuzhong Zhao, Yue Liu, Junpeng Liu, Jingye Chen, Xun Wu, Yaru Hao, Tengchao Lv, Shaohan Huang, Lei Cui, Qixiang Ye, Fang Wan, and Furu Wei. Geometric-Mean Policy Optimization. In *International Conference on Learning Representations*, pp. 41354–41374, 2026b.

Shuai Zhen, Yanhua Yu, Ruopei Guo, Nan Cheng, and Yang Deng. Hierarchical Reinforcement Learning with Augmented Step-Level Transitions for LLM Agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7035–7053, 2026.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group Sequence Policy Optimization. arXiv preprint, 2025a. arXiv:2507.18071.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. ProcessBench: Identifying Process Errors in Mathematical Reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1009–1024, 2025b.

Haizhong Zheng, Jiawei Zhao, and Beidi Chen. Prosperity before Collapse: How Far Can Off-Policy RL Reach with Stale Data on LLMs? In *International Conference on Learning Representations*, pp. 82657–82679, 2026a.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023.

Zhi Zheng, Rongsheng Chen, Yunpeng Ba, Zhenkun Wang, Yee Whye Teh, and Wee Sun Lee. Agentic ESOpt: Fine-Tuning Long-Horizon LLM Agents with Minimal GPU Requirements. arXiv preprint, 2026b. arXiv:2608.17310.

Tianle Zhong, Neiwen Ling, Yifan Pi, Zijun Wei, Tianshu Yu, Geoffrey Fox, Peng Wu, and Xiao Yu. Diagnosing Training Inference Mismatch in LLM Reinforcement Learning. arXiv preprint, 2026. arXiv:2605.14220.

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. ArCHer: Training Language Model Agents via Hierarchical Multi-Turn RL. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62178–62209, 2024.

Yuzhen Zhou, Jiajun Li, Yusheng Su, Gowtham Ramesh, Zilin Zhu, Xiang Long, Chenyang Zhao, Jin Pan, Xiaodong Yu, Ze Wang, Kangrui Du, Jialian Wu, Ximeng Sun, Jiang Liu, Qiaolin Yu, Hao Chen, Zicheng Liu, and Emad Barsoum. APRIL: Active Partial Rollouts in Reinforcement Learning to Tame Long-tail Generation. arXiv preprint, 2025. arXiv:2509.18521.

Zijian Zhou, Ao Qu, Zhaoxuan Wu, Sunghwan Kim, Alok Prakash, Daniela Rus, Bryan Kian Hsiang Low, and Paul Liang. MEM1: Learning to Synergize Memory and Reasoning for Efficient Long-Horizon Agents. In *International Conference on Learning Representations*, pp. 58413–58438, 2026.

Deyu Zou, Yongqiang Chen, Fan Feng, Mufei Li, Pan Li, Yu Gong, and James Cheng. On Information Self-Locking in Reinforcement Learning for Active Reasoning of LLM agents. In *Forty-third International Conference on Machine Learning*, 2026a.

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*, pp. 21511–21545, 2026b.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, Biqing Qi, Youbang Sun, Zhiyuan Ma, Lifan Yuan, Ning Ding, and Bowen Zhou. TTRL: Test-Time Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

APPENDIX CONTENTS

A. Formal Results and Boundary Conditions 21
B. Experimental Protocols and Results 23
C. Interface Literature Reconstruction 26
D. Interface Diagnostic Criteria 28
E. Interface Method Taxonomy 29
F. Cross-Interface Evidence Map 35

A FORMAL RESULTS AND BOUNDARY CONDITIONS

This appendix states the definitions, assumptions, and proofs used by the formal results in the main text. For a fixed Belief encoder ϕ and Granularity rule σ, p denotes the current-policy law of the random element ξ carrying one gradient term, q its behavior law, $w = \mathrm{d}p/\mathrm{d}q$ the exact density ratio, and $\bar{w}$ the implemented correction. We write $s = \nabla_\theta \log \pi_\theta(\bar{a} \mid c)$, let A be the current-policy advantage in the induced decision process, and write the implemented Signal as $\hat{A} = A + \epsilon$.

A.1 TASK-RELATIVE CONTROL SUFFICIENCY

Let $Q^\star(h, a)$ denote the optimal continuation value of action a after history h, and let $V^\star(h) = \sup_a Q^\star(h, a)$. Given a reference distribution ν over decision histories, define the task-relative control loss of $c = \phi(h)$ as

$$\mathcal{L}_\phi(\nu) = \inf_{\bar{\pi}} \mathbb{E}_{h\sim\nu}\left[V^\star(h) - \mathbb{E}_{a\sim\bar{\pi}(\cdot|\phi(h))} Q^\star(h, a)\right],$$

where $\bar{\pi}$ is restricted to policies that condition only on $\phi(h)$. We call ϕ δ-control-sufficient under ν when $\mathcal{L}_\phi(\nu) \le \delta$. This quantity is the irreducible one-decision regret imposed by the representation. Let ν_ϕ be the pushforward law of $c = \phi(h)$ and $\nu(\cdot \mid c)$ a regular conditional law of histories. For a finite action space with attained optima, the conditional loss is linear in the action distribution, giving

$$\mathcal{L}_\phi(\nu) = \mathbb{E}_{c\sim\nu_\phi}\left[\min_a \mathbb{E}_{h\sim\nu(\cdot|c)}[V^\star(h) - Q^\star(h, a)]\right].$$

Since $V^\star(h) - Q^\star(h, a) \ge 0$, $\mathcal{L}_\phi(\nu) = 0$ if and only if, for ν_ϕ-almost every c, there exists an action distribution supported on actions that are optimal for $\nu(\cdot \mid c)$-almost every history. If the essential intersection of these optimal-action sets is empty on a set of contexts with positive ν_ϕ-measure, then the control loss is strictly positive.

Approximate-information-state results bound sequential value loss when the representation approximately preserves immediate rewards and controlled transition predictions under suitable metrics (Subramanian et al., 2022).

A.2 TOKEN–DECISION SCORE EQUIVALENCE

Let I_k index the policy-owned tokens in a segment of the realized interaction record, with initial context c_k. Fix the ownership mask and segment boundaries when regrouping the record. The policy likelihood factor and its score are

$$L_{\theta,k} = \prod_{j\in I_k} \pi_\theta(x_j \mid x_{<j}, c_k),$$

$$\nabla_\theta \log L_{\theta,k} = \sum_{j\in I_k} \nabla_\theta \log \pi_\theta(x_j \mid x_{<j}, c_k).$$

The prefix $x_{<j}$ includes environment feedback observed before token j. For a complete policy-generated string with no intervening environment response, $L_{\theta,k} = \pi_\theta(\bar{a}_k \mid c_k)$.

If the tokens in I_k share the same advantage $\hat{A}_k$ and implemented correction $\bar{w}_k$, the corresponding updates also agree exactly:

$$\bar{w}_k \hat{A}_k \nabla_\theta \log L_{\theta,k} = \sum_{j\in I_k} \bar{w}_k \hat{A}_k \nabla_\theta \log \pi_\theta(x_j \mid x_{<j}, c_k).$$

With conditioning contexts and ownership masks fixed, token-specific advantages or correction weights can change the update without changing the score sum. If an additional boundary or termination variable is sampled, its own log-likelihood score must be included unless its realization is already encoded by a policy-owned token in I_k.

If a parser maps several complete policy-generated strings to the same semantic action a, then

$$\pi_\theta(a \mid c) = \sum_{x:\mathrm{parse}(x)=a} \pi_\theta(x \mid c).$$

The sum runs over disjoint complete realizations, including termination. For a sampled latent action, the likelihood is defined on the latent variable. Exact off-policy correction requires p and q to assign likelihood to the same joint event under a reconstructible context, boundary, and memory version.

A.3 Exact Off-Policy Correction: Support and Variance

When $p \ll q$, change of measure gives $\mathbb{E}_q[wX] = \mathbb{E}_p[X]$ for any integrable estimand X. If $\mathbb{E}_q[w^2\|X\|^2] < \infty$ (that is, $wX \in L^2(q)$), then

$$\mathbb{E}_q\Big[\|wX - \mathbb{E}_p X\|^2\Big] = \mathbb{E}_q[w^2\|X\|^2] - \|\mathbb{E}_p X\|^2.$$

Exact correction therefore removes expectation bias, while its variance depends on the joint tails of the density ratio and the estimand. For the constant scalar estimand $X \equiv m$,

$$\mathrm{Var}_q[wm] = m^2\big(\mathbb{E}_q[w^2] - 1\big) = m^2\chi^2(p\|q).$$

Cauchy–Schwarz gives

$$d_{\mathrm{TV}}(p,q) = \tfrac{1}{2}\mathbb{E}_q|w-1| \le \tfrac{1}{2}\sqrt{\mathbb{E}_q(w-1)^2} = \tfrac{1}{2}\sqrt{\chi^2(p\|q)}.$$

Replacing w with a clipped, truncated, or estimated correction $\bar{w}$ produces the exact expectation residual

$$\mathbb{E}_q[\bar{w}X] - \mathbb{E}_p[X] = \mathbb{E}_q[(\bar{w} - w)X].$$

Distributional distance measures the separation between behavior and target laws. $\bar{w} - w$ is the error in the implemented density-ratio correction.

A.4 Finite-Difference Expansion and Comparison Scope

For $x \in \{0,1\}^{\{B,G,S,R\}}$, let Δ_i set coordinate i to one minus setting it to zero, and define $D_U = (\Delta^U F)(\mathbf{0})$. Möbius inversion on the Boolean lattice gives

$$F(x) = \sum_{U \subseteq \mathrm{supp}(x)} D_U.$$

Evaluating at $x = \mathbf{1}$ and subtracting $D_\varnothing = F(\mathbf{0})$ proves Equation (4.1). Terms with $|U| = 1$ are main effects; the $\binom{4}{2} = 6$ terms with $|U| = 2$ are pairwise interactions; the remaining five terms collect third- and fourth-order interactions.

Each configuration corner must map to a common estimand. Signal and Rollout comparisons can preserve the induced decision representation directly. Belief or Granularity changes may also alter the policy class, conditioning context, or action space; such comparisons require a bridge through a shared task-level return, matched trajectories, or a deterministic regrouping of policy-owned events. The finite differences characterize non-additivity on this common response surface, while the shared operators and matched interventions identify its mechanism.

A.5 Proof of Theorem 5.1 and Boundary Conditions

Assume $p \ll q$ and that the terms below are integrable. Starting from $\tilde{g} = \mathbb{E}_q[\bar{w}s(A+\epsilon)]$ and $g = \mathbb{E}_p[sA]$, add and subtract $\mathbb{E}_q[ws(A+\epsilon)]$. Change of measure gives $\mathbb{E}_q[ws(A+\epsilon)] = \mathbb{E}_p[s(A+\epsilon)]$, hence

$$\tilde{g} - g = \mathbb{E}_p[s\epsilon] + \mathbb{E}_q[(\bar{w} - w)sA] + \mathbb{E}_q[(\bar{w} - w)s\epsilon].$$

Conditional on the realized policy, evaluator, and implementation state, p is the current-policy target law and q the law of samples used in the update. Exact correction sets $\bar{w} = w$ and removes both Rollout residuals in expectation. In the on-policy synchronous ideal, $p = q$ and $w = \bar{w} = 1$, leaving only $\mathbb{E}_p[s\epsilon]$; a nominally synchronous implementation may still have engine- or sampling-level mismatch. Summing over decision indices yields the trajectory-gradient identity when the corresponding terms are absolutely summable in expectation.

If $\|s\| \le M_s$, $|A| \le M_A$, and $|\epsilon| \le M_\epsilon$, then

$$\|\tilde{g} - g\| \le M_s \, \mathbb{E}_p|\epsilon| + M_s(M_A + M_\epsilon) \, \mathbb{E}_q|\bar{w} - w|.$$

The first term is defined under the current-policy distribution. If an evaluator is trained or audited under a distribution r_S and $p \ll r_S$, then, with $\rho_S = \mathrm{d}p/\mathrm{d}r_S$,

$$\mathbb{E}_p[s\epsilon] = \mathbb{E}_{r_S}[\rho_S s\epsilon].$$

Offline accuracy or calibration can therefore bound the realized gradient bias only when the audit distribution covers the current policy and the weighted quantity $\rho_S s\epsilon$ is controlled.

B Experimental Protocols and Results

B.1 Experimental Design

The study comprises four matched experimental blocks. Table 2 lists these blocks.

Block	Policy	Comparison	Updates	Evaluation
1.5B lag sweep	Qwen2.5-1.5B	Outcome / PRM / Judge × sync / lag 1, 2, 4, 8	500	3 seeds; 12 trajectories per checkpoint
1.5B controls	Qwen2.5-1.5B	MC-uniform / strong PRM × sync / lag 8	300	2–3 seeds; 12 trajectories per checkpoint
7B single-seed block	Qwen2.5-7B LoRA	Outcome / PRM / MC-dense × sync / lag 8	300	seed 42; fixed 100-task OOD evaluation set
Group-size-16 pair	Qwen2.5-1.5B	Outcome / PRM at synchrony	300	seed 42; 32-episode evaluations

Table 2: Experimental blocks. The 500-update lag sweep and 300-update controls are reported as separate campaigns.

Matched cells share the task pool, SFT initialization, optimizer, training budget, and evaluation protocol; they differ only in Signal construction or rollout lag. Across blocks, the two scales additionally differ in fine-tuning regime (full-parameter at 1.5B, LoRA at 7B), interaction horizon, PRM instantiation, and evaluation protocol; cross-block comparisons are therefore read as qualitative replications rather than matched contrasts. The Group-size-16 pair tests whether the synchronous Outcome–PRM direction persists under a larger sampling group. The corresponding results are reported in Table 5. ALFWorld supplies binary terminal success. Learned evaluators and outcome-derived Monte Carlo estimators provide alternative decision-level credit. The 7B PRM is trained on trajectories from disjoint tasks; its pre-optimization held-out AUROC is 0.860 with ECE 0.041. Complete configurations and implementation details are released with the code.

B.2 Evaluation Metrics

For each run in the 1.5B lag sweep and control round, the endpoint is the mean of its final three evaluations. These run-level endpoints are averaged across seeds. Checkpoint maxima are retained as secondary trajectory statistics, and the training run is the unit of inference. The 1.5B lag sweep and controls use 12 held-out evaluation trajectories every 25 updates, with 20 evaluations per 500-update run. The reward curve in Figure 2(d) reports the per-update mean PRM reward. The 7B tables report both the checkpoint mean and the final evaluation; for the 1.5B lag sweep and control round, the endpoint above is the primary quantity and checkpoint means serve only as summary trajectories.

For the logged action-conditional ratio w_{act}, define the behavior-context-averaged conditional shift and its estimator from n logged decision events within an update as

$$d_{\mathrm{TV}}^{\mathrm{act}} = \tfrac{1}{2}\mathbb{E}_q|w_{\mathrm{act}} - 1|, \qquad \widehat{d}_{\mathrm{TV}}^{\mathrm{act}} = \frac{1}{2n}\sum_{i=1}^{n}|w_{\mathrm{act}}^{(i)} - 1|,$$

together with clipping mass, ratio moments, and, for the 7B runs, the mean absolute ratio change from clipping. For the 1.5B PRM monitor, the evaluator target for a trajectory with K decision units is $y_k = (2\mathbf{1}\{\text{success}\} - 1)\gamma^{K-k-1}$ with $\gamma = 0.95$, matching PRM training; $\hat{r}_k$ denotes the evaluator score. Evaluator diagnostics include sign accuracy on collected rollout states, the empirical maximum score magnitude, and the mean score–label gap

$$g_{\mathrm{SL}} = \left|\mathbb{E}_{\mathcal{D}_R}[\hat{r} - y]\right|,$$

where $\mathcal{D}_R$ is the empirical decision-state sample collected by the corresponding rollout process.

B.3 Empirical Results

All six 7B seed-42 runs contain updates 0–299 and all six held-out evaluations. Table 3 reports the complete trajectories.

Signal	Rollout	50	100	150	200	250	300	ckpt. mean	final
Outcome	sync	.88	.97	.97	.99	.99	.99	.965	99/100
Outcome	lag 8	.90	.89	.92	.91	.95	.95	.920	95/100
PRM	sync	.87	.72	.00	.00	.00	.00	.265	0/100
PRM	lag 8	.89	.88	.88	.86	.89	.91	.885	91/100
MC-dense	sync	.89	.90	.99	1.00	.99	1.00	.962	100/100
MC-dense	lag 8	.90	.89	.96	.97	1.00	1.00	.953	100/100

Table 3: Qwen2.5-7B seed-42 confirmation with matched outcome-derived MC-dense controls. "Checkpoint mean" averages the six evaluation checkpoints; "final" reports the exact success count on the fixed 100-task OOD evaluation set. The common SFT baseline is 89/100.

At lag 8, policy-ratio TV, computed from ratios with recomputed behavior-policy probabilities in the denominator, averages 0.00574 for Outcome and 0.00713 for PRM over 300 updates. For ratios with generation-time behavior probabilities in the denominator, the mean absolute change from clipping averages 2.76×10^{-4} and 7.18×10^{-4}, respectively, over the same updates.

Table 4 reports training-trajectory diagnostics for the same four runs. $\mathrm{KL}_{\mathrm{ref}}$ is the per-token estimator $\exp(r) - r - 1$ with $r = \log \pi_{\mathrm{ref}} - \log \pi_\theta$, averaged over the policy-owned tokens of each update against the frozen SFT adapter, and is the penalty term of the training objective. Engine-ratio TV is $\frac{1}{2}\mathbb{E}|w_{\mathrm{eng}} - 1|$, where w_{eng} compares the behavior policy's likelihood recomputed by the training forward pass with the likelihood recorded by the generation engine for the same decision record. Gradient norm is the global norm before clipping. For PRM-sync, $\mathrm{KL}_{\mathrm{ref}}$ has a mean of 0.9 over updates 225–249 and a median of 0.11 over updates 150–299.

Signal / Rollout	updates	$\mathrm{KL}_{\mathrm{ref}}$	engine-ratio TV	grad. norm	train success
Outcome, sync	50–99	0.010	0.00093	1.05	0.81
	100–149	0.014	0.00088	0.93	0.87
	150–299	0.016	0.00087	0.96	0.90
Outcome, lag 8	50–99	0.010	0.00098	0.94	0.78
	100–149	0.010	0.00087	0.87	0.85
	150–299	0.018	0.00080	0.66	0.89
PRM, sync	50–99	0.029	0.00164	3.05	0.68
	100–149	0.117	0.00763	5.46	0.11
	150–299	0.258	0.00895	6.29	0.00
PRM, lag 8	50–99	0.017	0.00086	2.05	0.76
	100–149	0.040	0.00072	2.29	0.82
	150–299	0.045	0.00082	2.85	0.81

Table 4: Training-trajectory diagnostics of the 7B seed-42 block; each entry is the mean over the updates in the window. Gradient norms are averaged over finite values. Held-out trajectories for the same runs are in Table 3.

Signal	50	100	150	200	250	300	ckpt. mean	final
Outcome	.000	.000	.031	.063	.094	.125	.052	4/32
PRM	.000	.000	.000	.000	.000	.000	.000	0/32

Table 5: Qwen2.5-1.5B group-size-16 synchronous sensitivity pair. Both runs contain the complete 300-update training record and all six 32-episode evaluations; "final" reports the last evaluation, as in Table 3.

Table 6 summarizes the broader 1.5B campaign. The complete lag sweep is reported in Table 7.

Signal	provenance	updates	sync		lag 8	
			endpoint	max	endpoint	max
Outcome	environment	500	0.278	0.389	0.241	0.389
MC-uniform	outcome-derived	300	**0.602**	0.722	0.542	0.625
PRM (learned)	learned proxy	500	0.000	0.000	0.019	0.167
PRM (strong, 0.93)	learned proxy	300	0.000	0.042	0.028	0.250

Table 6: Endpoint summary. Endpoint is the mean of the last three evaluations; max is the mean of each run's maximum. Outcome and learned PRM belong to the 500-update lag sweep; MC-uniform and strong-PRM runs belong to the 300-update control round. Outcome, learned PRM, and MC-uniform sync use seeds $\{42, 7, 1024\}$; strong PRM uses $\{42, 7\}$; MC-uniform lag 8 uses $\{42, 1024\}$. The three Outcome sync runs are at $0.250/0.167/0.167$ (seeds $42/7/1024$) at their first evaluation. The strong PRM's 0.93 denotes its sign accuracy.

Signal	provenance	sync	lag 1	lag 2	lag 4	lag 8
Outcome	environment	0.278	0.296	0.259	0.259	0.241
Learned PRM	learned proxy	0.000	0.000	0.000	0.000	0.019
Frozen Judge	learned proxy	0.000	0.000	0.000	0.019	0.028

Table 7: Endpoint success across the complete 500-update lag sweep. Each cell averages three seeds after computing the final-three-evaluation mean within each run.

step	runs	PRM acc. on $\mathcal{D}_R$	g_{SL}	$\max_{\mathcal{D}_R} \lvert\hat{r}\rvert$
0	15	0.835	0.103	0.980
100	15	0.903	0.181	0.984
200	15	0.564	0.534	0.988
300	15	0.446	0.821	0.992
400	15	0.257	1.087	0.996

Table 8: Evaluator monitor across the 15 main PRM runs. Entries are checkpoint-wise means over runs; g_{SL} is the mean score–label gap and $\max_{\mathcal{D}_R} |\hat{r}|$ is the empirical maximum score magnitude.

Table 8 reports the evaluator monitor. The bounded score magnitude remains near its tanh limit throughout. Across the fifteen main PRM runs, accuracy on collected rollout states at the last evaluator checkpoint ranges from 0 to 0.125 under synchrony and lags 1–2 (nine runs, mean 0.054) and from 0.2375 to 0.8125 under lags 4 and 8 (six runs, mean 0.560), with no overlap between the bands; a Jonckheere–Terpstra trend test across the five lag levels gives one-sided $p = 0.02$ (10^6 Monte Carlo permutations), and the mean score–label gap is smaller under lags 4 and 8. The mean score–label gap changes with the behavior process and is not the score-weighted target-law quantity $\mathbb{E}_p[s\epsilon]$ in Equation (5.1); it compares the mean evaluator score with the mean outcome-derived label on those states. Figure 3 reports the corresponding evaluator trajectories across rollout lags.

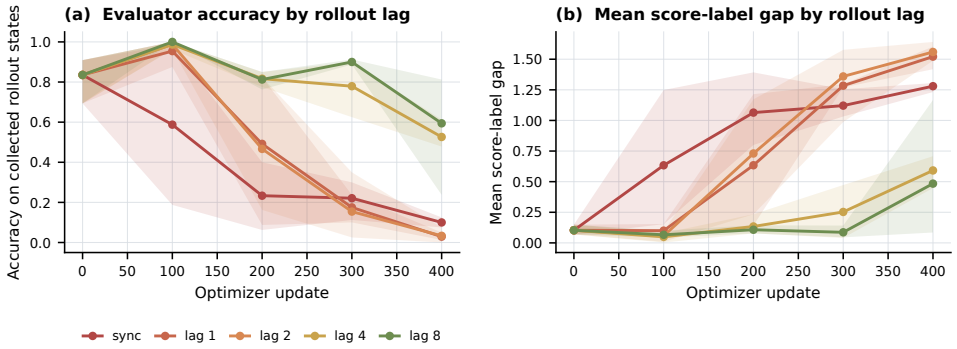

Figure 3: Evaluator behavior on collected rollout states, stratified by rollout lag, across the fifteen main 1.5B PRM runs (three seeds per lag). Lines are means over seeds and bands span the seed range. **(a)** Accuracy on states supplied by each rollout process. **(b)** Mean score–label gap.

C Interface Literature Reconstruction

C.1 Belief: Control-Relevant State Construction

Belief methods differ in where information is stored, how it is represented, and how it is updated. Managed in-context methods build the state inside the model input through window selection and running summaries (Xiao et al., 2024; Wu et al., 2026). MEM1 learns to coordinate memory and reasoning (Zhou et al., 2026). Long-context studies show that evidence recoverable from the context may still fail to change behavior, separating information presence from policy use (Liu et al., 2024; Gao et al., 2024). External non-parametric memory stores evidence outside the active context and retrieves fragments on demand (Lewis et al., 2020; Packer et al., 2024; Rasmussen et al., 2025; Xu et al., 2025b), at the cost of a longer construction chain in which storage, retrieval, and reintegration can introduce omission, ranking, and staleness errors. Learned memory brings that chain under policy control, either by making memory operations actions optimized from long-term return (Yan et al., 2026b; Yu et al., 2026b) or by maintaining a parametric module of approximately constant size (Wang et al., 2024c; 2025b; Ren et al., 2026); the information state then depends on a memory version, so identical parameters no longer imply identical behavioral conditions, and experience reuse must identify the state under which an action was generated. Explicit-belief methods make the semantics of the state testable by representing hypotheses, evidence provenance, and uncertainty (Lidayan et al., 2026; Singh et al., 2026; Liao et al., 2026a), which guards against premature commitment to a single account under partial observability but shifts the burden onto update calibration.

Belief methods are therefore compared by task-relative control loss (Appendix A.1) together with the retrieval, update, and inference cost of constructing the state. Counterfactual interventions on memory content test whether retained evidence changes action rankings as predicted (Appendix D).

C.2 Granularity: Policy Ownership and Decision Factorization

Granularity methods address policy ownership, decision boundaries, and temporal abstraction. Ownership masking excludes environment tokens from the policy loss in trajectories that concatenate model responses with execution results, delimiting the events that enter the score (Jiang et al., 2026a; Xue et al., 2026). External feedback nonetheless shifts the conditional distribution of subsequent generation, an effect that belongs to Belief and Rollout rather than to masking. Execution protocols define semantic boundaries at the level of turns, tool calls, programs, GUI operations, and embodied actions (Wei et al., 2026; Liang et al., 2023; Dong et al., 2026; Luo et al., 2025a; Qi et al., 2025). They can also represent each model invocation as a transition with its own input, output, and reward (Luo et al., 2025b). Cross-scale methods keep the relation between timescales explicit, from the options framework (Sutton et al., 1999) to turn-level critics (Zhou et al., 2024), Bellman backups with action decomposition (Wen et al., 2024), and coupled hierarchical advantages at subgoal boundaries (Peng et al., 2026). ARIA aggregates returns over semantically clustered history–action pairs (Yang

et al., 2025). GiGPO groups recurring anchor states to estimate local relative advantages (Feng et al., 2025). Both retain action boundaries and modify the Signal operator.

Comparisons among Granularity choices rest on four contracts: policy ownership of the decision unit, consistent reconstruction of its probability from generation, assignment of credit to the same event, and definition of behavior correction on the same joint random variable. The separating test changes σ on fixed trajectories and identifies the partitions that actually move the gradient, the action ranking, or a resolvable counterfactual consequence.

C.3 Signal: Objective Anchoring and Credit Construction

Environment rewards, program tests, and outcome verifiers anchor credit directly in task success (Le et al., 2022; Lambert et al., 2025; Guo et al., 2025a). Monte Carlo methods estimate continuation values from rollouts sampled at intermediate states (Kazemnejad et al., 2025; Guo et al., 2025b). SPA-RL uses a learned progress estimator to attribute terminal reward to individual steps (Wang et al., 2025a). Under the undiscounted episodic objective, potential-based shaping with a fixed potential, $r'_k = r^\star_k + \Phi(c_{k+1}) - \Phi(c_k)$, preserves optimal policies if the potential is zero at every episode endpoint (Ng et al., 1999). Verifiable process rewards localize external checks to individual events where intermediate correctness is testable (Yuan et al., 2026).

Learned evaluators amortize that evidence instead of re-invoking it: a value critic estimates continuation return, a PRM scores intermediate steps, and judges or rubrics score partial or complete trajectories against criteria (Lightman et al., 2024; Xi et al., 2026; Liu et al., 2026a). Policy updates change the state–action visitation distribution. Value critics need to track current-policy continuation returns. Process and trajectory evaluators need to maintain agreement with task outcomes on states visited by the policy (Gao et al., 2023; Zheng et al., 2025b; Song et al., 2025; Zhang et al., 2025d). PRIME learns implicit process rewards from current-policy rollouts and outcome labels, iStar updates step rewards from trajectory preferences, and APRM extends error coverage with progressively harder adversarial negatives (Cui et al., 2026; Liu et al., 2026b; Juneja et al., 2026). Aggregation choices such as sum versus minimum over the same PRM outputs change sensitivity to local error and compensation across steps (Cheng et al., 2025).

MOPD and SDPO construct dense advantages from teacher–student log-probability differences on student-generated trajectories (Ma et al., 2026; Hübotter et al., 2026). SDAR combines task RL with a gated self-distillation loss using privileged context (Lu et al., 2026). Reference information can also be converted into scalar turn rewards, temporal-difference credit, or token weights (Tao et al., 2026; Shan et al., 2026). Privileged re-scoring compares the policy's own token log probabilities across standard and privileged contexts to construct credit signals (Wang et al., 2026h; Yang et al., 2026b). OCSD isolates the contribution of future observations by contrasting structurally matched replay views (Yang et al., 2026b). The aggregation rule determines the decision unit that receives this token-level evidence (Wang et al., 2026h).

Evidence can finally be derived from the model itself. Pseudo-labels from majority outcomes, training signals from semantic entropy or output concentration, and self-certainty estimated from the model distribution all dispense with external supervision (Zuo et al., 2025; Shafayat et al., 2025; Zhang et al., 2025b; Prabhudesai et al., 2026; Zhao et al., 2026a). In prolonged self-rewarded training, majority-vote rewards can increase even as task accuracy collapses (Shafayat et al., 2025). Self-generated curricula additionally bring the task distribution into the learning process: Absolute Zero generates and solves problems while a code executor validates well-formedness and answer correctness, retaining anchoring in executable semantics (Zhao et al., 2025), and R-Zero co-evolves a Challenger and Solver so that difficulty is regulated endogenously (Huang et al., 2026a). In self-generated curricula, evaluation rules determine credit at the Signal interface, while task generators alter the distribution of initial contexts.

Signal constructions divide further by their effect on the optimum. A chosen objective folded into per-decision reward enters through $\hat{A}_k$ and produces an intentional deviation relative to the fixed task objective; baselines and valid potential shaping alter credit without changing the task optimum, while a separate loss adds its own gradient term. Signal methods are therefore distinguished by evidence provenance and calibration distribution rather than feedback density. Relevant measurements include sign accuracy and score–label gap on an identified target- or behavior-policy distri-

bution, process–outcome agreement, and task response when density is held fixed while provenance changes.

C.4 Rollout: Behavior Distribution

Synchronous on-policy optimization keeps behavior likelihood, context, and execution state traceable across a short update cycle, while limiting stale-data reuse (Schulman et al., 2017; Shao et al., 2024). As inference cost and interaction horizon grow, waiting for the slowest rollout makes strict synchronization a systems bottleneck. Importance sampling with exact density ratios estimates target-policy expectations from behavior samples. IMPALA's V-trace uses truncated per-decision ratios to mitigate the effects of actor–learner policy lag (Espeholt et al., 2018). For a unit composed of policy-owned token indices I_k, with μ_R the behavior policy's conditional law, the ratio factorizes as

$$\frac{\pi_\theta(\bar{a}_k \mid c_k)}{\mu_R(\bar{a}_k \mid c_k)} = \prod_{j \in I_k} \frac{\pi_\theta(x_j \mid x_{<j}, c_k)}{\mu_R(x_j \mid x_{<j}, c_k)}.$$

Rollout and Granularity are coupled through the decision unit: token, prefix, turn, and sequence ratios correct different random events, so the aggregation boundary must agree with the unit used by the policy score and advantage (Zheng et al., 2025a; Li et al., 2026a). Autoregressive ratios accumulate with decision span and policy divergence, so a few extreme values can dominate gradients on long trajectories. Weight tapering modifies sample weights, and geometric aggregation changes how weighted terms are combined (Le Roux et al., 2025; Zhao et al., 2026b). M^{2}PO uses a second-moment criterion on log importance ratios to mask token updates (Zheng et al., 2026a). GAC uses gradient projection to control persistently aligned stale gradients (Xu et al., 2026). Filtering by age, ratio, or support instead changes the effective q_R, trading reuse for a narrower training distribution. Asynchronous systems decouple generation from learning and balance workloads while managing sample staleness (Noukhovitch et al., 2025; Fu et al., 2025). Partial-rollout continuation carries unfinished responses across policy updates, allowing a trajectory to span multiple policy versions (Zhou et al., 2025).

Subsequent contexts depend on tool outputs, environment branches, retrieved information, and memory updates, so identical checkpoints can induce different context occupancy; a local action ratio then corrects only the conditional action law at a context, and reweighting the paths that produce that context requires prefix or trajectory correction. Training and inference engines may assign different probabilities to the same tokens under identical weights and input (Zhong et al., 2026), and sampling temperature, nucleus truncation, and search-guided or best-of-n generation have the same character: they separate the behavior policy from π_θ at zero lag. Recording the sampling rule is necessary for correction; exact recovery additionally requires preservation of target support or a target policy defined with the same truncation or search rule. When tools, memory, or inference implementations change version, action ratios alone cannot recover the joint probability of the contexts they generated.

Rollout methods can be compared along two dimensions: the grain at which correction is applied, and how closely the implemented $\bar{w}$ tracks the exact ratio on the events that carry credit. Total variation, ratio moments, clipping mass, and $\mathbb{E}_q|\bar{w} - w|$ characterize distribution mismatch and correction residuals. Correction granularity determines which behavior likelihoods and generating contexts must be reconstructed for reused trajectories.

D Interface Diagnostic Criteria

Table 9 operationalizes the four interface contracts through matched diagnostics. The controlled study instantiates the Signal and Rollout criteria; the Belief and Granularity entries specify corresponding comparison designs.

Interface	Diagnostics
Belief	counterfactual intervention on memory content; predictive sufficiency for subsequent rewards and transitions; retrieval recall and freshness; evidence provenance; uncertainty calibration; ability to revise an erroneous state
Granularity	policy ownership of the unit; reconstructibility of unit probability from generation; agreement of the credit event with the scored event; definition of behavior correction on the same joint random variable; sensitivity of gradient, action ranking, and resolvable counterfactuals to a change of σ on fixed trajectories
Signal	sign accuracy and score–label gap on an identified target- or behavior-policy distribution; agreement between process scores and terminal outcomes; response of task success when feedback density is held fixed while provenance changes
Rollout	policy-ratio total variation; ratio moments; clipping mass; residual $\mathbb{E}_q\lvert\bar{w} - w\rvert$; reconstructibility of behavior likelihood and generating context for every reused trajectory

Table 9: Diagnostics associated with each interface contract.

E Interface Method Taxonomy

This appendix organizes the method families discussed in §3 and §4 by the primary mathematical object they change, and compares their mechanisms, operating conditions, characteristic failures, and cross-interface dependencies. Signal$\times$Rollout admits the exact residual identity of Theorem 5.1 for a fixed decision process; the remaining pairings correspond to the state-dependent, operator-matching, and conditional mechanisms analyzed in §4.

E.1 Review Scope and Coding

The review uses a framework-driven representative corpus whose unit of inclusion is a technically specified mechanism that changes an object in Equation (2.2) or supplies evidence about an interface contract. It combines foundational results on POMDPs, policy gradients, credit assignment, and off-policy learning with the survey literature cited in Section 1 and recent agentic RL methods; representative methods are selected to span distinct mechanisms and operating assumptions. Each family is coded by the primary interface object it directly perturbs and by the dominant structural pairing or pairings on which its validity depends.

E.2 Signal

Tables 10 and 11 compare Signal families by provenance, anchoring, and credit operator. This separates four technically different ways to obtain dense learning signals: verify intermediate states, estimate value from outcomes, learn an evaluator, or project a teacher policy. "Dense" therefore describes the placement of feedback, not its semantics.

Family	Representative methods	Mechanism / perturbed object	Characteristic failure	Primary coupling exposure(s)
External anchors				
Verifiable outcome rewards	(Le et al., 2022; Lambert et al., 2025; Guo et al., 2025a)	program, test, or environment scores the outcome within specified coverage	sparse credit; blind spots or noisy verifiers (Cai et al., 2026; Pignatelli et al., 2024)	reference anchor for all three other interfaces
Outcome-derived and objectively verifiable process signals				
Outcome-derived credit estimation	(Kazemnejad et al., 2025; Guo et al., 2025b; Yang et al., 2025; Zhang et al., 2025c)	derives credit or preferences from sampled outcomes or semantic history–action groups, with shortest-path reward estimation in ReasonRAG	sampling error or value differences obscured by aggregation	Signal×Rollout + Signal×Belief
Verifiable process rewards	(Yuan et al., 2026)	checks intermediate states with symbolic, algorithmic, or environment-grounded rules	unavailable when intermediate correctness is not objectively testable	Signal×Granularity
Learned evaluators and value proxies				
Return redistribution	(Wang et al., 2025a)	uses a learned progress estimator to assign stepwise credit	target depends on boundary and redistribution operator	Signal×Granularity
Learned step-PRMs	(Lightman et al., 2024; Wang et al., 2024a; Luo et al., 2024; Yuan et al., 2025; Liu et al., 2026b; Juneja et al., 2026)	learns step-level reward scores from process labels or trajectory-level supervision	reward hacking; limited generalization to harder problems (Zheng et al., 2025b); process errors missed despite correct answers (Song et al., 2025; Zhang et al., 2025d)	Signal×Rollout (Thm. 5.1)
Agentic process evaluators	(Xi et al., 2026; Chae et al., 2025; Zhang et al., 2026c)	scores agent actions from interaction context or full trajectories with reference answers	evaluation error under distribution shift	Signal×Rollout + Signal×Belief
LLM judges and critique models	(Zheng et al., 2023; McAleese et al., 2024; Yuan et al., 2024)	evaluates responses through scores or textual critiques	biased judgments or incorrect critiques	Signal×Belief
Rubric / generative RMs	(Liu et al., 2026a)	decomposes outcome into synthetic criteria for non-verifiable domains	judge biases exploited (Wang et al., 2026e)	Signal×Belief
Learned value functions	(Zhou et al., 2024; Yue et al., 2025a)	learns value estimates from returns or TD targets for advantage estimation	value-estimation bias and unstable critic learning	Signal×Rollout

Table 10: Signal strategies grounded in external outcomes, verifiable processes, or learned evaluators.

Family	Representative methods	Mechanism / perturbed object	Characteristic failure	Primary coupling exposure(s)
Teacher-policy projection on student states				
Multi-teacher on-policy distillation	(Ma et al., 2026)	teacher–student log-probability differences become token-level advantages on student trajectories	inherits teacher support, routing, and calibration limits	Signal×Rollout + Signal×Belief
Feedback-conditioned self-distillation	(Hübotter et al., 2026)	feedback-conditioned self-teacher supplies logit-level advantages	uninformative or misleading feedback; weak in-context learning	Signal×Belief
Reference-derived and self-conditioned credit				
Reference / self credit	(Tao et al., 2026; Shan et al., 2026)	turn/state TD rewards or token weights from a frozen reference model or self-conditioning	reference-answer sensitivity; multiplicative weights can distort scale	Signal×Granularity + Signal×Belief
Privileged re-scoring	(Wang et al., 2026h; Yang et al., 2026b)	log-probability contrasts between standard and privileged contexts	replay scaffold shifts scores (OCSD); aggregation rule sets the decisions identified as pivotal (AgentOPSD)	Signal×Belief + Signal×Granularity
Hybrid and endogenous objectives				
Additive hybrids	(Feng et al., 2025; Zhang et al., 2026d; Xu et al., 2025a; Wei et al., 2026)	local or dense credit combined with exogenous outcome	destabilizes without dimensional calibration (Xu et al., 2025a)	Signal×Granularity
Belief-quality credit	(Liu et al., 2026c; Tang et al., 2026)	combines task rewards with belief-entropy or belief–observation consistency rewards	belief scores misaligned with task return	Signal×Belief
Model-internal rewards	(Zuo et al., 2025; Zhang et al., 2025b; Zhao et al., 2026a; Yuan et al., 2024; Wu et al., 2025b)	self-consistency, confidence, or self-judging supplies endogenous training targets	gains that reflect amplified pretrained behavior rather than label correctness (Shao et al., 2026); reward hacking in prolonged self-rewarded training (Shafayat et al., 2025)	Signal×Belief; Signal×Rollout
Self-generated curricula	(Zhao et al., 2025; Huang et al., 2026a; Liu et al., 2025)	task generation changes the initial-context distribution; an evaluator separately assigns credit	curriculum and evaluator can co-evolve toward a self-consistent but task-misaligned fixed point	Task-distribution axis; Signal through the evaluation rule
Policy-invariant shaping under explicit conditions				
Potential-based shaping	(Ng et al., 1999; Xie et al., 2026)	adds potential differences consistent with the task return	inconsistent return convention or episode-endpoint handling	Signal×Belief

Table 11: Teacher-policy projection, reference-derived credit, hybrid objectives, endogenous rewards, and invariant shaping.

E.3 BELIEF

Table 12 compares Belief families by information locus, representation semantics, and update control (§3.1). Representational capacity is evaluated relative to task-relevant distinctions and control loss.

Family	Representative methods	Mechanism / perturbed object	Characteristic failure	Primary coupling exposure(s)
Managed in-context state				
Window truncation	(Xiao et al., 2024)	retains KV states of initial attention-sink tokens and recent tokens	may remove action-relevant evidence	Signal×Belief + Belief×Granularity
Context summary	(Park et al., 2023; Shi et al., 2026; Wu et al., 2026; Kang et al., 2026; Zhou et al., 2026; Zhen et al., 2026)	compresses history into a running state	lossy or goal-misaligned state; accessible information may still be unused (Liu et al., 2024; Gao et al., 2024)	Signal×Belief + Belief×Granularity
Self-reflective update	(Shinn et al., 2023; Madaan et al., 2023; Gou et al., 2024)	adds feedback or reflections to the context to guide subsequent generation	missed errors or new errors introduced during revision	Signal×Belief + Belief×Granularity
External non-parametric state				
Retrieval	(Lewis et al., 2020; Packer et al., 2024; Wang et al., 2026g)	assembles context from an external index	retrieval omissions and ranking errors	Signal×Belief + Belief×Granularity
Structured memory	(Rasmussen et al., 2025; Xu et al., 2025b)	temporal knowledge graph or atomic-note store	heuristic construction may omit or distort control-relevant relations	Signal×Belief + Belief×Granularity
Structured scratch-pads	(Yao et al., 2023; Press et al., 2023)	writable buffer for intermediate reasoning	re-import still bounded by window	Belief×Granularity (writes as split points)
Action-controlled, explicit, and parametric state				
Memory-as-action	(Yu et al., 2026a; Wang et al., 2025c; Yan et al., 2026b; Yu et al., 2026b)	read, write, or edit is an explicit policy action	competes with task actions under sparse credit	Belief×Granularity + Belief×Rollout
Parametric memory	(Wang et al., 2024c; 2025b; Ren et al., 2026)	fixed-size latent or weights update online	memory state can drift with both policy and replay distribution	Belief×Rollout when memory versions differ between collection and update
Explicit belief modules	(Lidayan et al., 2026; Jiang et al., 2026b; Singh et al., 2026; Liao et al., 2026a)	natural-language or compact uncertainty-bearing state is decoupled from policy text	calibration and evidence provenance require separate evaluation (Liu et al., 2026c)	Signal×Belief + Belief×Rollout (Zou et al., 2026b)

Table 12: Belief strategies by information locus, state semantics, and update control (§3.1).

E.4 GRANULARITY

Table 13 compares token-level, semantic, and learned decision units (§3.2). Credit and correction operators must use compatible factorizations of the decision event defined by σ.

Family	Representative methods	Mechanism / perturbed object	Characteristic failure	Primary coupling exposure(s)
Policy-owned versus externally produced tokens				
Output / tool-token masking	(Jiang et al., 2026a; Xue et al., 2026)	excludes environment-produced tokens from the policy score	environment tokens treated as policy outputs during optimization	Granularity ×Rollout
Schema-defined semantic decisions				
Turn / interaction boundaries	(Wei et al., 2026; Li et al., 2026b; Xi et al., 2025)	uses interaction turns as policy decision units and applies trajectory-level advantages	terminal credit remains coarse within a turn	Signal×Granularity + Granularity ×Rollout
Hierarchy-based segments	(Zhou et al., 2024; Peng et al., 2026; Gao et al., 2026a)	lifts the unit to multi-step macro-actions	segmentation declared a priori (HiPER learns the switch)	Signal×Granularity
Subgoal decomposition	(Prasad et al., 2024; Wang et al., 2026d)	decomposes the task into subgoal units	subgoals can be marked complete before the required action sequence ends	Signal×Granularity
Code / command actions	(Liang et al., 2023; Wang et al., 2024b; Mai et al., 2025; Yang et al., 2024)	expresses actions as executable code or interface commands	coarse, execution-dependent credit	Signal×Granularity + Granularity ×Rollout
Tool-integrated reasoning	(Patil et al., 2024; Du et al., 2024; Jin et al., 2025; Dong et al., 2026; Wang et al., 2026b)	the tool call is the action boundary	boundary tied to tool schema	Signal×Granularity + Granularity ×Rollout
GUI / embodied schemas	(Luo et al., 2025a; Qi et al., 2025)	unit set at the operation or trajectory level	domain-fixed unit	Signal×Granularity + Granularity ×Rollout
Expanded action space	(Yue et al., 2025b)	adds routing and environment-specific actions beyond the vocabulary	rarely sampled routing actions limit training on new actions	Signal×Granularity + Granularity ×Rollout

Table 13: Granularity families from token to learned abstraction (§3.2).

E.5 Rollout

Table 14 separates correction *grain* from correction *tightness*. The former specifies the token, prefix, segment, or sequence used to construct correction weights. The latter controls estimator tails, residual bias, and data reuse. Off-policy correction uses the behavior probabilities and conditioning inputs associated with sampled actions.

Family	Representative methods	Mechanism / perturbed object	Characteristic failure	Primary coupling exposure(s)
Algorithm layer: importance-ratio grain				
Token-level ratio	(Zhao et al., 2026b)	finest correction grain; geometric aggregation of token-level ratios	variance over long responses; MoE instability	Granularity ×Rollout
Turn / subsentence / prefix ratio	(Li et al., 2026a; Yang et al., 2026a; Lei et al., 2026)	geometric-mean turn/subsentence ratios, or current token ratios scaled by the smallest earlier token ratio (MinPRO)	boundary mismatch and length-sensitive variance	Granularity ×Rollout
Sequence-level ratio	(Zheng et al., 2025a; Huang et al., 2026b)	length-normalized sequence ratios (GSPO), truncated sequence ratios (VCPO), and token–sequence mixing (Min et al., 2026)	sequence weights concentrated on a few responses	Granularity ×Rollout (grain match)
Algorithm layer: grain-orthogonal tightness controls				
Clipping / divergence control	(Yu et al., 2025; Le Roux et al., 2025)	bounds the ratio at the chosen unit	residual post-clipping bias	Signal×Rollout (Thm. 5.1)
Stale-data stabilization	(Zheng et al., 2026a; Xu et al., 2026)	applies log-ratio-based token masking (M^2PO) or projects persistently aligned stale gradients (GAC)	high-variance weights or stale-aligned gradient collapse	Signal×Rollout; GAC acts in gradient space
Stale-data correction	(Espeholt et al., 2018; Zheng et al., 2026a)	uses V-trace correction or log-ratio-based token masking	residual bias from weight truncation or token masking	Signal×Rollout
Partial rollout	(Kimi Team et al., 2026; Zhou et al., 2025; Qu et al., 2025)	pauses unfinished trajectories and resumes them in later iterations	behavior mismatch when resumed under an updated policy	Signal×Rollout
Sync baseline	(Schulman et al., 2017; Shao et al., 2024)	alternates rollout collection with repeated updates on each batch	throughput bottleneck	Signal×Rollout
System layer: rollout execution				
Async frameworks	(Noukhovitch et al., 2025; Fu et al., 2025; Wu et al., 2025a)	decouples rollout generation from policy updates	stale samples from earlier policy versions	Signal×Rollout (Thm. 5.1)
Rollout infrastructure	(Sheng et al., 2025; Cao et al., 2025; Gao et al., 2025; Lu et al., 2025; Zhang et al., 2026b)	schedules rollout execution and data delivery to the trainer	policy lag when collection overlaps policy updates	Signal×Rollout
Train–inference mismatch	(Zhong et al., 2026)	different token probabilities under identical weights and inputs	behavior–target mismatch even at zero lag	Signal×Rollout

Table 14: Rollout families by grain and tightness of the behavior–target correction (§3.4).

F Cross-Interface Evidence Map

Figure 4 organizes the six structural pairings by their shared object. Table 15 summarizes representative phenomena, relevant literature, measurable diagnostics, and formal status for the same interface pairs.

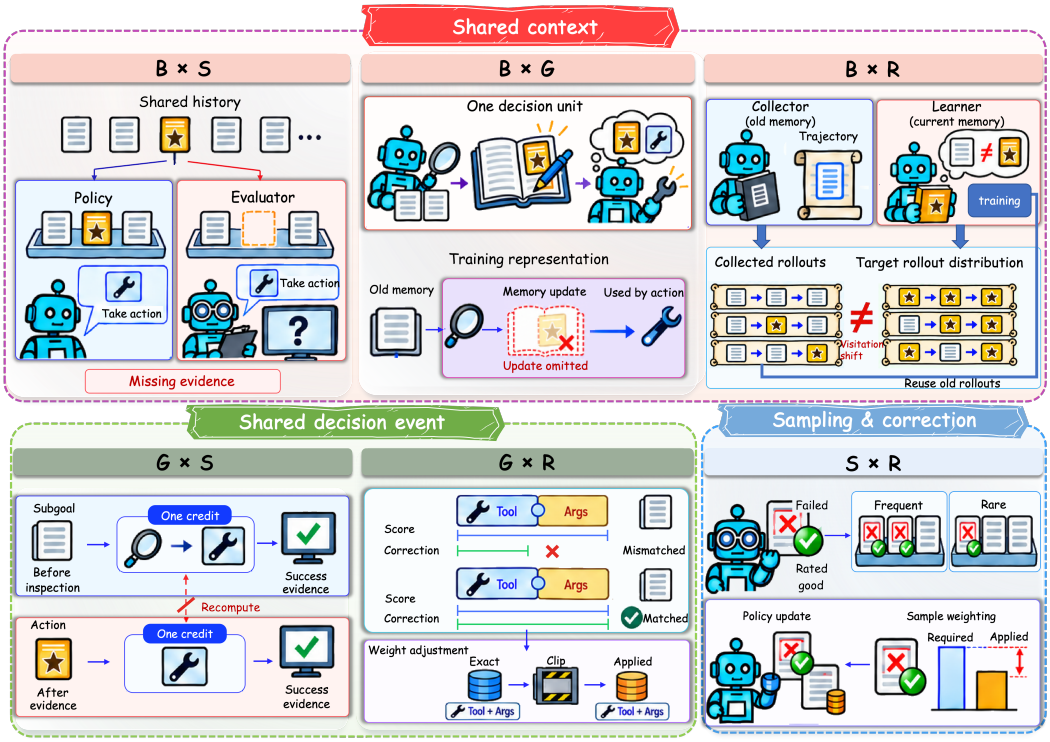

Figure 4: Six structural pairings among the four interfaces, grouped by their shared conditioning context c_B, decision event $\bar{a}_G$, or behavior measure $(q_R, \bar{w}_R)$.

Table 15: Cross-interface evidence map for the six interface pairs. This table lists evidence for each pairing; the coupling-exposure column of Appendix E instead marks which pairing a method family is exposed to.

Coupling	Representative phenomenon	Literature	Measurable diagnostic	Formal status
Signal × Rollout	Learned proxies and outcome-derived credit can respond differently to the same rollout shift; reward-model optimization can widen proxy–outcome divergence	(Gao et al., 2023; Zhang et al., 2025d; Cui et al., 2026; Juneja et al., 2026; Le Roux et al., 2025; Zheng et al., 2026a; Xu et al., 2026)	Current-policy evaluator error and score alignment; $d_{\mathrm{TV}}(p, q)$; clipping mass; $\mathbb{E}_q\lvert\bar{w} - w\rvert$	Exact estimator identity for a fixed decision process (Thm. 5.1)
Signal × Belief	Missing or miscalibrated historical information changes both action selection and evaluator judgment	(Zou et al., 2026b;a; Tang et al., 2026; Singh et al., 2026; Yang et al., 2026b)	Action ranking, evaluator calibration, and control loss under memory interventions	Conditional-expectation projection identity (Eq. (4.2))

Coupling	Representative phenomenon	Literature	Measurable diagnostic	Formal status
Signal × Gran.	The same feedback induces different credit operators at token, segment, turn, and trajectory boundaries	(Guo et al., 2025b; Wang et al., 2026a; Gao et al., 2026b; Peng et al., 2026; Wang et al., 2026f;h)	Advantages, policy scores, and outcome agreement recomputed under alternative boundaries	Operator constraints through ownership, credit, correction, and counterfactual resolution
Gran. × Rollout	Importance-ratio grain and policy-score events jointly determine asynchronous variance and correction bias	(Zheng et al., 2025a; Zhao et al., 2026b; Li et al., 2026a; Xue et al., 2026)	Ratio moments, clipping mass, gradient variance, and event-level likelihood reconstruction	The density ratio must be defined over the random event in the score-weighted estimand
Belief × Rollout	Changes in memory version or behavior context make action-only correction incomplete	(Ren et al., 2026; Yan et al., 2026a; Liao et al., 2026b; Singh et al., 2026)	Memory-version mismatch, context reconstruction, stale-gradient cosine, and rejection rate	Version-aware correction requirement for mutable state and behavior context
Belief × Gran.	Subgoal, memory-action, and provenance boundaries aligned with belief transitions change hierarchical credit	(Peng et al., 2026; Wang et al., 2026d; Li et al., 2026c; Liao et al., 2026b; Yan et al., 2026a)	Boundary–memory alignment, predictive sufficiency at split points, and hierarchical-advantage variance	Empirical characterization: belief-aligned versus fixed-boundary comparisons