BEYOND TIMESTAMPS: DECISION-ALIGNED ON-POLICY DISTILLATION FOR LONG-HORIZON AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards (RLVR) often relies on sparse outcome rewards, providing coarse supervision for long-horizon agents. On-policy self-distillation (OPSD) complements this signal with dense privileged feedback. However, we identify *Decision–Timestamp Mismatch*: privileged guidance may be misaligned with the student's functional decision because the corresponding decision can occur at a different timestep, while the student's decision itself may span multiple timesteps rather than being tied to a single timestamp. Thus, timestamp-local supervision can misalign both the context and the temporal scope of credit. To address this mismatch, we introduce ALIGNOPSD, following the principle of aligning supervision before assigning credit. Decision-Aligned Supervision Rectification re-scores the same student-sampled response in functionally matched contexts across sibling rollouts to calibrate local teacher evidence. Semi-Markov Hierarchical Credit Assignment then derives variable-duration decision spans from correspondence changes and uses rectified evidence to allocate outcome-grounded credit across spans and their constituent turns. We evaluate ALIGNOPSD with Qwen2.5-3B and Qwen2.5-7B on ALFWorld, WebShop, and Search-QA against representative baselines. ALIGNOPSD outperforms both GRPO and StepOPSD across all eight backbone–aggregate-metric comparisons, improving on GRPO by 5.5–8.7 % and ranking first in six. Additional analyzes examine the two alignment stages and hyperparameter sensitivity between tasks.

1 INTRODUCTION

Agentic long-horizon tasks require language agents to accomplish goals through a sequence of interdependent intermediate decisions, while training rewards are often available only at the end of a task. Methods such as Group Relative Policy Optimization (GRPO) compare terminal rewards across sibling rollouts and broadcast a trajectory-level advantage to the generated tokens (Shao et al., 2024). Such supervision captures overall task performance, but cannot directly distinguish effective decisions from incidental behaviors within successful trajectories, nor can it precisely localize errors within failed ones (Feng et al., 2025; Zhang et al., 2026b). To complement this coarse-grained supervision, on-policy distillation (OPD) and its self-distilled variants (OPSD) provide dense teacher feedback on behaviors generated by the student policy itself (Agarwal et al., 2024; Zhao et al., 2026). When the teacher is further granted privileged task information, such feedback can provide richer local evidence about intermediate behaviors (Penaloza et al., 2026; Lu et al., 2026). The challenge is therefore not merely to obtain denser supervision, but to determine which teacher evidence can meaningfully inform the student's intermediate decisions at different stages of execution.

Toward more effective intermediate supervision, recent agentic OPSD methods have developed along two main directions. One line enriches teacher evidence through hindsight information, execution feedback, or privileged contexts (Lu et al., 2026; Wang et al., 2026b; Yu et al., 2026; Yang et al., 2026b; Zhang et al., 2026a), while another refines how teacher–student discrepancies are filtered and aggregated into action- or turn-level learning signals (Zhang et al., 2026c; Li et al., 2026a). These advances improve the informativeness and utilization of supervision, yet their local signals remain anchored to corresponding positions along the student trajectory. Interpreting these local discrepancies as decision-level credit relies on a key assumption: *teacher evidence at the current trajectory position provides an appropriate basis for evaluating the functional decision pursued by*

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

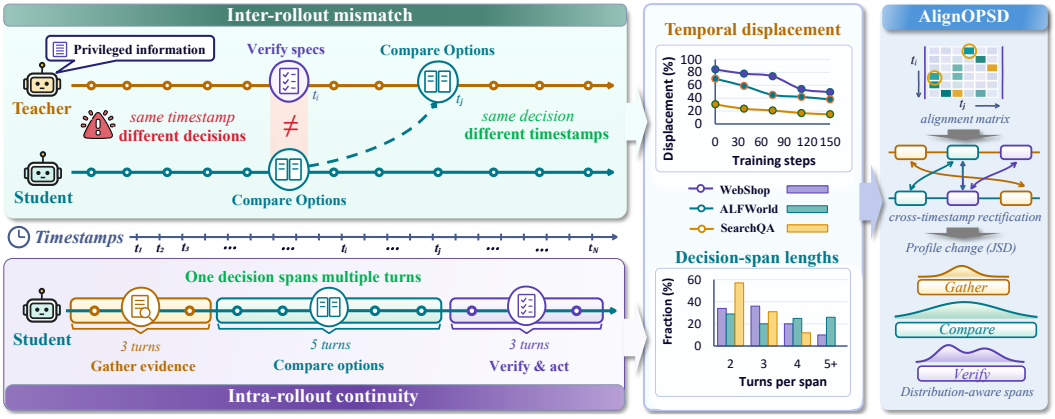


Figure 1: **Decision–Timestamp Mismatch.** Functional decisions need not share timestamps across different sibling rollouts and can span multiple turns within individual trajectories.

the student at that position. However, these refinements do not guarantee functional correspondence between teacher and student behaviors across rollouts, timestamps, and decision boundaries.

We refer to this phenomenon as *Decision–Timestamp Mismatch*, which manifests at two related levels. Across rollouts, the same timestamp need not identify the same decision context, while corresponding decisions may occur at different turns. As illustrated in Figure 1, the student compares product options while the privileged teacher verifies specifications, the corresponding comparison appears later in the teacher rollout. Even under the same visible history, privileged information may change which decision the teacher favors. Consequently, although same-prefix probability comparisons remain valid, the resulting disagreement need not isolate the quality of the student's current decision. The issue is therefore the functional comparability of supervisory contexts for reliable decision-level credit attribution, not the validity of same-prefix probability comparisons.

Within a rollout, a functional decision may persist across multiple turns, with boundaries that do not coincide with individual timestamps. Assigning credit independently to each turn can fragment behaviors that jointly realize one decision, whereas fixed-length windows can mix distinct decisions. The upper-right panels illustrate how these two aspects can be characterized quantitatively through the temporal displacement of corresponding decisions and the distribution of decision-span lengths. Motivated by this, we study: ***How can functional correspondence guide both the selection of privileged evidence and the temporal scope of outcome-grounded credit assignment?***

To address this problem, we propose ALIGNOPSD, which aligns privileged supervision with functional decisions before assigning credit. Specifically, Decision-Aligned Supervision Rectification estimates functional correspondences across sibling rollouts, re-scores the same student-sampled response in matched privileged contexts, and combines cross-context and local evidence according to correspondence confidence. Building on this alignment, Semi-Markov Hierarchical Credit Assignment constructs variable-duration decision spans from correspondence changes, allocates outcome-grounded credit across spans and their constituent turns using rectified evidence, and broadcasts turn-level credit to the corresponding training tokens. We evaluate ALIGNOPSD on ALFWorld, WebShop, and Search-QA under matched training and evaluation protocols against representative agentic RL methods spanning embodied, search, and shopping environments.

Our contributions are summarized as follows:

- **Problem Identification.** We identify *Decision–Timestamp Mismatch*: timestamp-local privileged evidence may not correspond to the student's functional decision, while the same decision can occur at positions across rollouts, leading to mismatched supervisory contexts and credit horizons.
- **Proposed Solution.** We propose ALIGNOPSD, which establishes functional correspondence before credit assignment by rectifying privileged supervision with matched contexts and deriving adaptive decision spans for outcome-grounded credit allocation over variable-duration decisions.
- **Experimental Validation.** Across three agentic benchmarks, ALIGNOPSD improves over GRPO by 5.5–8.7 % in all eight backbone–metric comparisons and ranks first in six. Additional analyses characterize correspondence alignment, span adaptation, and supervision rectification.

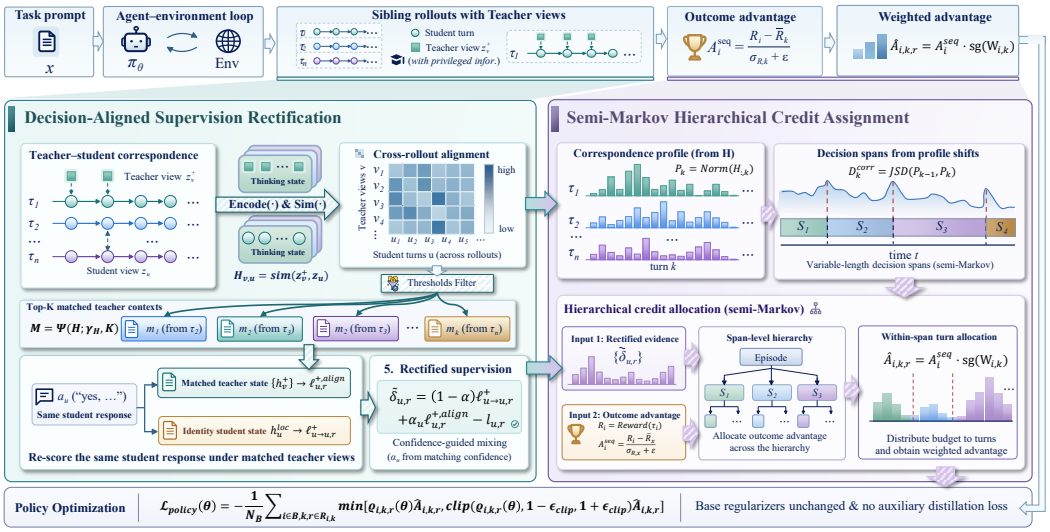


Figure 2: **Overview of ALIGNOPSD.** Cross-rollout alignment establishes comparable supervisory contexts, while correspondence changes define decision spans for hierarchical credit assignment. The weights enter policy optimization, with auxiliary components used during training.

2 METHODOLOGY

2.1 PROBLEM FORMULATION

Given task x and initial observation o_0, trajectory i starts from $h_{i,1} = (x, o_0)$. At turn k,

$$a_{i,k} = (y_{i,k,r})_{r=1}^{L_{i,k}} \sim \pi_\theta(\cdot \mid h_{i,k}), \qquad h_{i,k+1} = h_{i,k} \oplus (a_{i,k}, o_{i,k}), \tag{1}$$

where $a_{i,k}$ may contain a thinking trace followed by an executable action, $o_{i,k}$ is the observation returned by the environment, and $\oplus$ appends the response–observation pair to the interaction history. After K_i turns, trajectory τ_i receives a verifiable terminal reward $R_i = R(\tau_i)$. For each task x, group-relative policy optimization samples G sibling trajectories $\mathcal{G}_x$ and computes

$$A_i^{\text{seq}} = \frac{R_i - \overline{R}_x}{\widehat{\sigma}_{R,x} + \epsilon}, \qquad \overline{R}_x = \frac{1}{G} \sum_{j \in \mathcal{G}_x} R_j. \tag{2}$$

The base objective is broadcast A_i^{seq} to the optimized response tokens of the trajectory i, providing a scale and direction update based on the results. In OPSD-style training, each interaction history $h_{i,k}$ is paired with a privileged view $h_{i,k}^+ = (h_{i,k}, k_x)$, where k_x is task-relevant privileged information obtained through relevance-based retrieval and available only during training. Both views share the same policy π_θ. We denote the ordinary student view by $\pi_\theta(\cdot \mid h_{i,k})$ and define the privileged teacher view as $\pi_\theta^+(\cdot \mid h_{i,k}) \triangleq \pi_\theta(\cdot \mid h_{i,k}^+)$. Thus, the teacher differs from the student only in its conditioning context, rather than its model parameters, as shown in Figure 2.

2.2 DECISION-ALIGNED SUPERVISION RECTIFICATION

For compact notation, let $u = (i, k)$ denote a target turn and $v = (j, l)$ a candidate source turn from another sibling rollout. Conventional OPSD evaluates a_u only under its timestamp-local privileged view h_u^+, which may be misaligned with the functional decision represented at u. We therefore establish cross-rollout decision correspondence before rectifying the local teacher evidence.

Decision-state correspondence. Decision-relevant information may be distributed across observations and interaction history, making direct state matching difficult. We therefore use thinking traces as semantic surrogates of the decision state. Let z_u denote the thinking parsed from student response a_u, and z_v^+ the privileged thinking generated by the teacher view $\pi_\theta^+(\cdot \mid h_v)$. A frozen semantic encoder $\mathbf{Enc}(\cdot)$ maps both into a shared decision-state space:

$$\mathbf{d}_u = \mathbf{Enc}(z_u), \qquad \mathbf{d}_v^+ = \mathbf{Enc}(z_v^+). \tag{3}$$

We treat proximity in this space as approximate functional-decision correspondence. Appendix D.2 audits this surrogate with a method-faithful comparison of exact same-state and cross-rollout pairs.

Decision-alignment operator. Given the student and teacher-side decision-state representations, we construct a source-by-target correspondence matrix

$$H_{v,u}^{(n)} = \cos\Big(\mathbf{d}_v^{+,(n)}, \mathbf{d}_u^{(n)}\Big), \qquad \mathcal{M}^{(n)} = \Psi\Big(H^{(n)}; \gamma_H, K\Big). \tag{4}$$

For each target turn u, Ψ considers structurally valid source turns from other rollouts of the same task, removes candidates below similarity threshold γ_H, retains at most one source per sibling rollout, and selects the *top-K* matches. Across all three benchmarks, an action-consistency filter requires valid parsed actions and an action-embedding cosine of at least 0.8. On WebShop, where actions admit exact symbolic checks, we additionally require matching operation types and identical click targets. Each nonzero column $\mathcal{M}_{\cdot,u}^{(n)}$ assigns normalized weights to the retained reference views, while unmatched targets receive a zero column. The operator is recomputed for each on-policy batch, allowing the correspondence structure to evolve with the policy during training.

Multi-view supervision rectification. Instead of corresponding to the identity view $v = u$, our alignment operator additionally provides off-diagonal reference views for the same target response. Rather than imitating retrieved responses, the privileged policy π_θ^+ teacher-forces the same complete response a_u under each aligned source history from a sibling rollout:

$$\ell_{v\to u,r}^{+} = \log \pi_\theta^{+}(y_{u,r} \mid h_v, y_{u,<r}), \qquad \ell_{u,r} = \log \pi_\theta(y_{u,r} \mid h_u, y_{u,<r}). \tag{5}$$

The identity view gives the conventional local gap $\delta_{u,r}^{\mathrm{id}} = \ell_{u\to u,r}^{+} - \ell_{u,r}$. For a matched target, $\mathcal{M}_{\cdot,u}^{(n)}$ aggregates the aligned privileged views and rectifies the local evidence as

$$\ell_{u,r}^{+,\mathrm{align}} = \log \sum_v \mathcal{M}_{v,u}^{(n)} \exp\big(\ell_{v\to u,r}^{+}\big), \quad \widetilde{\delta}_{u,r} = (1-\alpha_u)\ell_{u\to u,r}^{+} + \alpha_u \ell_{u,r}^{+,\mathrm{align}} - \ell_{u,r}. \tag{6}$$

The coefficient $\alpha_u \in [0, \alpha_{\max}]$ adaptively controls how strongly aligned views modify the local signal according to the quality of the match. Concretely, we compute

$$\widehat{H}_{v,u} = \mathrm{clip}\left(\frac{H_{v,u} - \gamma_H}{1-\gamma_H}, 0, 1\right), \qquad \rho_u = \sum_v \mathcal{M}_{v,u}^{(n)} \widehat{H}_{v,u}, \qquad \alpha_u = \alpha_{\max}\rho_u \tag{7}$$

If no valid source is retrieved, we set $\alpha_u = 0$, reducing $\widetilde{\delta}_{u,r}$ to the conventional identity gap $\delta_{u,r}^{\mathrm{id}}$. The rectified gap $\widetilde{\delta}_{u,r}$ is then passed to the hierarchical credit allocator in Section 2.3.

2.3 Semi-Markov Hierarchical Credit Assignment

Rectification provides decision-relevant supervisory references, but a functional decision may span multiple interaction turns. Thus, we organize credit over variable-duration decision spans and introduce a hierarchical allocator Φ that converts rectified evidence into outcome-grounded advantages.

Correspondence-driven spans. A continuing functional decision is expected to retain similar reference contexts across adjacent turns. We normalize correspondence scores over structurally valid sources from other same-task rollouts before top-K truncation:

$$\mathbf{P}_{i,k}^{(n)} = \mathrm{Norm}\Big(H_{\cdot,(i,k)}^{(n)}\Big), \qquad D_{i,k}^{\mathrm{corr}} = \mathrm{JSD}\Big(\mathbf{P}_{i,k-1}^{(n)}, \mathbf{P}_{i,k}^{(n)}\Big). \tag{8}$$

where $D_{i,k}^{\mathrm{corr}}$ measures the correspondence shift between adjacent turns. A batch-adaptive threshold and duration constraints produce a contiguous partition $\mathcal{S}_i = \{S_{i,1}, \ldots, S_{i,J_i}\}$, whose variable-duration spans serve as credit units under a semi-Markov abstraction.

Outcome-oriented evidence. The rectified gap provides token-level teacher support, while hierarchical allocation operates at the turn level. We therefore define an outcome-oriented evidence score measuring agreement between rectified support and the trajectory-level outcome direction:

$$\mathcal{E}_{i,k} = \mathrm{sgn}(A_i^{\mathrm{seq}}) \frac{1}{N_{i,k}} \sum_{r=1}^{N_{i,k}} \widetilde{\delta}_{i,k,r} \tag{9}$$

where $N_{i,k}$ denote the number of optimized response tokens at turn k, a larger $\mathcal{E}_{i,k}$ indicates stronger outcome-consistent teacher support and therefore stronger evidence for allocating credit to turn k. For span $S_{i,m}$, we define $\mathcal{E}^{\text{span}}_{i,m}$ as the mean evidence over all turns in the corresponding span.

Hierarchical credit allocation. We view A^{seq}_i as an outcome-grounded credit budget and use the decision hierarchy to redistribute it. Let $\mathbf{E}_i = (\mathcal{E}_{i,k})_k$ denote the turn-level evidence, and define the cumulative credit measure $\Lambda_{i,k} = \sum_{t \le k} N_{i,t}$, whose increments recover the uniform per-token allocation of the objective. We introduce a credit allocator Φ to produce turn-level credit weights:

$$\mathbf{W}_i = \Phi(\mathcal{S}_i, \mathbf{E}_i; \Lambda_i)\,. \tag{10}$$

Internally, Φ performs two-level allocation. For each span $S_{i,m}$, its aggregate evidence $\mathcal{E}^{\text{span}}_{i,m}$ determines a span-level budget $B_{i,m}$. Conditioned on this budget, turn-level evidence determines the within-span allocation $Q_{i,k|m}$:

$$B_{i,m} = \Phi_{\text{span}}\big(\mathcal{E}^{\text{span}}_{i,m}, \Delta\Lambda_{i,m}\big)\,, \qquad W_{i,k} = \text{Norm}[B_{i,m}\,\Phi_{\text{turn}}(\mathcal{E}_{i,k}, \Delta\Lambda_{i,k} \mid S_{i,m})]\,. \tag{11}$$

Thus, the outer allocation assigns span-level credit, while the inner allocation distributes it across constituent turns. After bounded normalization, turn weights reweight the trajectory advantage:

$$\widehat{A}_{i,k} = A^{\text{seq}}_i\,\text{sg}(W_{i,k}), \tag{12}$$

where sg denotes stop-gradient. KL regularization constrains both allocation levels from deviating excessively from the neutral credit measure, while normalization preserves the token-weighted mean advantage. Exact allocation and normalization details are provided in Appendix F.

Policy optimization. Each optimized token inherits its turn advantage, $\widehat{A}_{i,k,r} = \widehat{A}_{i,k}$. For the current trajectory batch $\mathcal{B}$, let $\pi_{\theta_{\text{old}}}$ denote the fixed snapshot used for rollout collection and evidence construction. Define the policy ratio and total optimized-token count as

$$\varrho_{i,k,r}(\theta) = \frac{\pi_\theta(y_{i,k,r} \mid h_{i,k}, y_{i,k,<r})}{\pi_{\theta_{\text{old}}}(y_{i,k,r} \mid h_{i,k}, y_{i,k,<r})}, \qquad N_{\mathcal{B}} = \sum_{i\in\mathcal{B}}\sum_{k} N_{i,k}. \tag{13}$$

The token-mean clipped policy loss is

$$\mathcal{L}_{\text{policy}}(\theta) = -\frac{1}{N_{\mathcal{B}}} \sum_{\substack{i\in\mathcal{B},\,k \\ r\in\mathcal{R}_{i,k}}} \min\Big[\varrho_{i,k,r}(\theta)\widehat{A}_{i,k,r}, \text{clip}(\varrho_{i,k,r}(\theta), 1-\epsilon_{\text{clip}}, 1+\epsilon_{\text{clip}})\,\widehat{A}_{i,k,r}\Big], \tag{14}$$

where ϵ_{clip} is the base clipping threshold. Other base-objective regularization terms remain unchanged, and no auxiliary distillation loss is added. All correspondence, span, evidence, and allocation quantities are stop-gradient and used only during training.

3 Experiments

3.1 Experimental Setup

Benchmarks. We evaluate on ALFWorld (Shridhar et al., 2021), Search-QA (Jin et al., 2025), and WebShop (Yao et al., 2022), covering embodied household control, search-augmented question answering, and interactive online shopping, respectively. We report success rate for ALFWorld, exact-match accuracy for Search-QA, and normalized score and exact success for WebShop. Dataset composition, splits, and per-category evaluation details are provided in Appendix E.

Baselines. We compare three groups: training-free prompting methods (Vanilla and Skill-Prompt*), RLVR methods (GRPO, Skill-GRPO, and Skill-GRPO*), and self-distillation RL methods (OPSD, GRPO+OPSD, Skill-SD, RLSD, SDAR, and StepOPSD) (Shao et al., 2024; Zhao et al., 2026; Wang et al., 2026a; Yang et al., 2026a; Lu et al., 2026; Zhang et al., 2026c). All skill-conditioned methods use the same SkillBank and retrieval source. Detailed configurations are given in Appendix F.1.

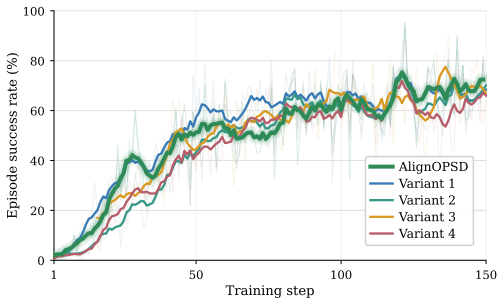
Implementation details. We use Qwen2.5-3B-Instruct and Qwen2.5-7B-Instruct backbones (Qwen et al., 2025). Training uses a single node with up to eight NVIDIA A800 GPUs. Rollout settings, hyperparameters, and training diagnostics are reported in Table 8 and Appendices F–G.

Table 1: **Performance on ALFWorld, Search-QA, and WebShop.** We report success rate (%) on ALFWorld, accuracy (%) on Search-QA, and Score/Acc (%) on WebShop. Skills are training-only unless marked with $*$ (validation with skills). **Best** and second-best are highlighted.

	ALFWorld							Search-QA								WebShop	
Method	**Pick**	**Heat**	**Look**	**Clean**	**Cool**	**Pick2**	**Avg**	**NQ**	**Triv**	**Pop**	**Hotp**	**2Wk**	**MuS**	**Bam**	**Avg**	**Score**	**Acc**
Qwen2.5-3B-Instruct																	
Vanilla	44.4	15.4	11.1	6.2	28.6	12.5	21.9	24.6	48.1	31.0	26.3	25.3	7.2	59.7	31.7	6.7	0.8
Skill-Prompt*	51.7	0.0	66.7	48.4	4.3	10.0	28.9	23.7	46.2	30.6	24.4	22.1	7.5	12.5	23.9	1.2	0.8
OPSD	48.8	0.0	41.7	16.7	15.8	16.7	28.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	11.3	3.1
GRPO	91.2	61.9	62.5	96.2	65.0	47.4	75.0	39.3	60.6	41.1	37.4	34.6	15.4	26.4	36.4	79.8	63.3
Skill-GRPO	88.9	70.6	71.4	58.8	40.7	29.2	60.2	43.5	58.8	43.0	36.8	32.2	11.7	12.5	34.1	77.3	60.9
Skill-GRPO*	94.3	66.7	57.1	**100**	73.1	57.1	80.5	44.3	59.6	44.3	39.0	36.1	14.5	14.9	36.1	76.3	66.4
GRPO+OPSD	**100**	75.0	**82.4**	85.7	70.0	60.0	**81.2**	44.3	60.8	46.0	38.9	38.6	13.9	64.5	43.9	77.8	66.4
Skill-SD	88.2	52.4	50.0	96.2	65.0	57.9	73.4	42.8	60.1	42.2	**39.5**	38.9	13.6	64.9	44.0	75.9	64.0
RLSD	87.9	75.0	75.0	90.9	73.1	68.4	79.7	42.3	58.6	43.3	**39.5**	**39.6**	**16.1**	**66.1**	44.2	84.4	66.4
SDAR	94.6	50.0	66.7	68.2	60.0	**80.0**	74.2	**44.8**	58.1	44.3	38.6	36.2	15.7	**66.1**	43.4	70.7	57.8
StepOPSD	82.4	52.2	66.7	82.6	73.7	75.0	73.4	43.6	**61.2**	43.8	39.2	38.1	15.8	64.5	43.7	82.4	66.4
ALIGNOPSD	88.6	**81.8**	54.5	90.3	**82.6**	58.8	80.5	44.5	**61.2**	**46.3**	**39.5**	38.9	13.7	65.3	**45.1**	**86.8**	**68.8**
Qwen2.5-7B-Instruct																	
Vanilla	36.1	0.0	22.2	3.1	0.0	0.0	12.5	25.2	50.8	29.5	29.0	29.0	10.4	63.7	33.9	5.9	1.6
Skill-Prompt*	51.7	5.3	50.0	32.3	4.3	0.0	23.4	30.9	52.1	32.7	32.7	27.9	12.7	66.1	36.4	1.7	0.8
OPSD	50.0	21.4	60.0	22.7	17.6	9.5	32.8	8.8	8.6	17.5	2.5	4.2	0.5	1.2	6.2	4.5	2.3
GRPO	91.2	81.0	87.5	96.2	65.0	57.9	81.2	45.1	63.7	44.0	43.6	43.2	16.8	37.6	42.0	80.9	72.6
Skill-GRPO	88.5	61.1	66.7	65.2	57.7	73.1	69.5	45.2	63.7	45.7	43.1	43.3	19.6	21.4	40.3	80.4	71.9
Skill-GRPO*	**100**	83.3	83.3	96.4	75.0	**78.9**	88.4	44.8	63.0	45.1	43.7	43.7	20.5	71.4	47.5	87.0	**81.2**
GRPO+OPSD	91.4	87.5	61.5	**100**	76.5	52.2	80.4	**47.3**	64.5	46.9	43.8	39.3	18.0	69.4	47.0	86.8	76.5
Skill-SD	93.9	**100**	**93.8**	90.9	69.2	68.4	85.1	47.1	64.5	47.8	44.2	42.1	20.2	69.0	47.8	86.1	76.5
RLSD	**100**	58.8	87.5	92.3	**80.0**	65.2	82.0	46.8	63.0	44.4	**45.5**	**48.9**	**21.5**	**73.0**	49.0	87.4	77.3
SDAR	94.7	86.7	75.0	**100**	68.2	**78.9**	85.9	46.3	63.5	48.2	43.8	48.4	19.6	**73.0**	49.0	83.6	74.2
StepOPSD	98.1	90.5	75.0	**100**	**80.0**	63.2	88.4	45.3	64.6	45.1	44.5	44.4	19.3	69.8	48.2	87.2	78.1
ALIGNOPSD	70.8	**100**	81.8	**100**	76.5	70.8	**89.1**	46.4	**65.6**	**48.5**	44.6	43.7	19.5	69.8	**49.1**	**87.9**	78.9

Table 2: **Ablation Study.** Final Score and Acc for ALIGNOPSD and four ablation variants are reported on the left and success-rate curves over 150 training steps are shown on the right.

#	Variant	Score	Acc
–	**ALIGNOPSD**	**87.9**	**78.9**
1	w/o rectification + adaptive allocation	84.2	71.9
2	w/ rectification + token allocation	81.6	71.1
3	w/ rectification + turn allocation	86.3	77.3
4	w/ rectification + random allocation	78.9	69.5



3.2 MAIN RESULTS

Aggregate performance. Table 1 compares ALIGNOPSD with trajectory-level RLVR and OPSD+RLVR baselines under the matched evaluation protocol. ALIGNOPSD achieves 80.5%/89.1% average success on ALFWorld, 45.1%/49.1% accuracy on Search-QA, and 86.8%/87.9% Score on WebShop for the 3B/7B backbones, respectively, under both model scales. Across the eight backbone–aggregate-metric comparisons, ALIGNOPSD ranks first in six and second in two, surpassed only by GRPO+OPSD on 3B ALFWorld (81.2%) and Skill-GRPO* on 7B WebShop Acc (81.2%).

Improvement over trajectory- and step-level credit. Compared with GRPO, ALIGNOPSD improves all eight aggregate comparisons, by 5.5%/7.9% on ALFWorld, 8.7%/7.1% on Search-QA, 7.0%/7.0% on WebShop Score, and 5.5%/6.3% on WebShop Acc for the 3B/7B backbones. ALIGNOPSD also outperforms StepOPSD across all eight comparisons, showing that the gains extend beyond step-level supervision alone. Since skill-conditioned baselines share the same source, the results suggest that privileged information alone is insufficient. Behavioral alignment and outcome-conditioned credit assignment also matter. We next ablate the two alignment stages.

3.3 ABLATION STUDY

Effect of supervision rectification. Table 2 reports WebShop results with the 7B backbone. Removing Decision-Aligned Supervision Rectification while retaining adaptive allocation reduces Score

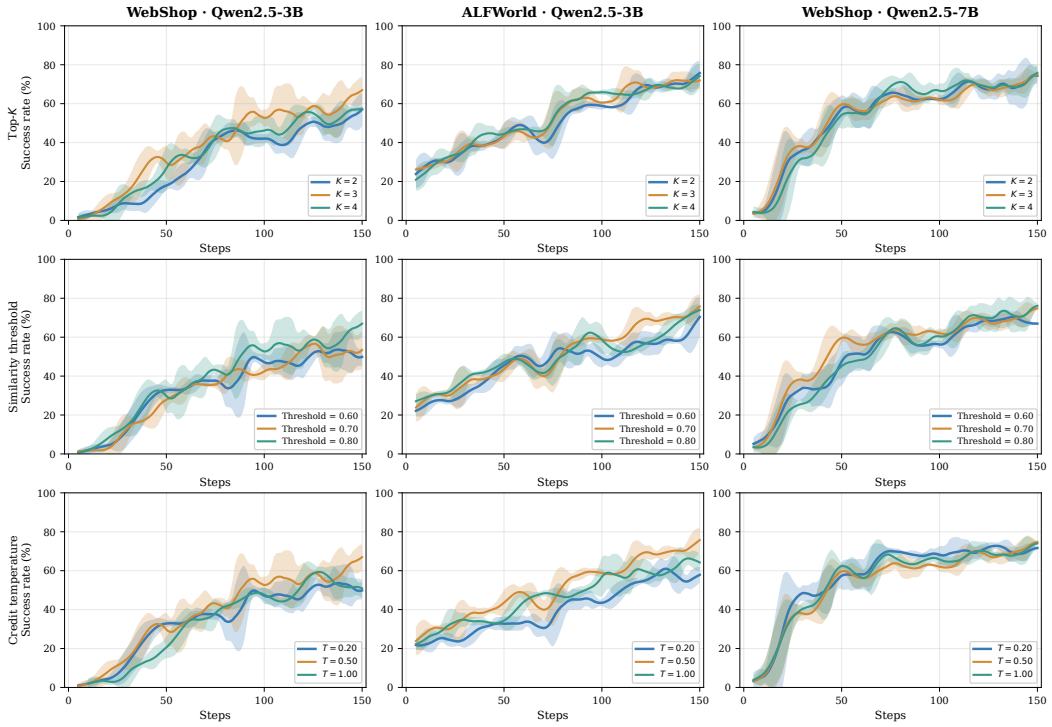


Figure 3: **Sensitivity Analysis.** Sensitivity to retrieval top-K, thinking-similarity threshold γ_H, and credit temperature T across the tested hyperparameter ranges. Curves smooth recorded validation checkpoints, and shading shows twice the local temporal variation rather than across-seed uncertainty.

from 87.9% to 84.2% and Acc from 78.9% to 71.9%, corresponding to drops of 3.7 and 7.0%. This confirms that establishing a comparable supervisory context contributes beyond the allocation mechanism. The larger degradation in Acc indicates that rectification is particularly important for converting local teacher–student discrepancies into credit that supports complete task success.

Effect of adaptive allocation. Keeping rectification fixed, we replace adaptive decision spans with token-, turn-, or random span-level allocation. Across both metrics, turn-level allocation outperforms token-level allocation, which in turn outperforms random allocation. Relative to these alternatives, ALIGNOPSD improves Score/Acc by 6.3/7.8% over token-level allocation, 1.6/1.6% over turn-level allocation, and 9.0/9.4% over random allocation. These results show that turns already provide meaningful units for credit assignment, whereas uniformly finer token-level credit does not yield better supervision. The additional gain over turn-level allocation suggests that decision boundaries should adapt to correspondence changes during policy optimization rather than remain fixed.

3.4 SENSITIVITY ANALYSIS

Figure 3 examines sensitivity to retrieval top-K, thinking-similarity threshold γ_H, and credit temperature T, which control evidence coverage, correspondence filtering, and credit concentration, respectively. All panels report validation success rate from development histories, the curves characterize empirical operating regions rather than across-seed uncertainty.

Retrieval top-K. The retrieval top-K controls retrieval breadth, balancing broader evidence coverage against the inclusion of less relevant matches. We evaluate $K \in \{2, 3, 4\}$. On ALFWorld 3B, the corresponding endpoints are 78.1%, 73.4%, and 76.6%, favoring $K = 2$, whereas WebShop 3B reaches 60.2%, 69.5%, and 57.8%, favoring $K = 3$. WebShop 7B instead favors $K = 4$, showing that no single value dominates across settings. Sensitivity also varies: the endpoint spread is only 4.7% on ALFWorld 3B and 4.6% on WebShop 7B, but increases to 11.7% on WebShop 3B. Thus, retrieval top-K exhibits a task- and scale-dependent coverage–selectivity trade-off rather than a monotonic trend across the tested benchmarks and backbone scales in practice.

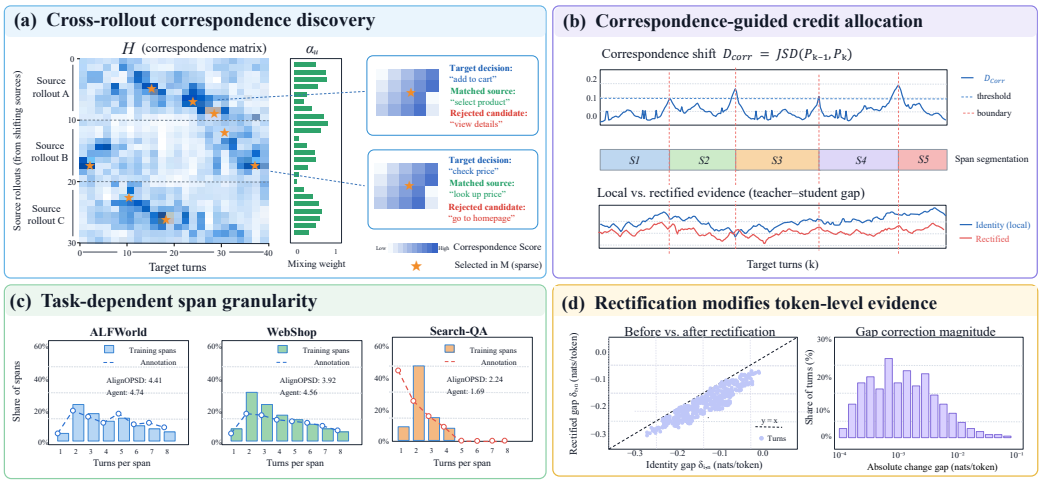


Figure 4: **Mechanistic diagnostics.** (a) Cross-rollout correspondence and confidence-weighted mixing; (b) correspondence-guided span segmentation and local versus rectified gaps; (c) task-dependent span-length distributions; (d) turn-averaged teacher–student gaps before and after rectification (left) and the distribution of absolute gap changes across target turns (right).

Thinking-similarity threshold γ_H. The thinking-similarity threshold γ_H filters weak correspondences before credit allocation, trading evidence retention against match quality. We evaluate $\gamma_H \in \{0.60, 0.70, 0.80\}$. WebShop 3B improves markedly from 51.6% to 69.5% as the threshold increases, while ALFWorld 3B peaks at $\gamma_H = 0.70$ with 78.1%. WebShop 7B similarly favors a higher threshold, reaching 76.6% at $\gamma_H = 0.80$. The endpoint spread is only 3.9% on ALFWorld 3B, compared with 17.9% on WebShop 3B and 9.4% on WebShop 7B. This indicates that correspondence filtering matters more for WebShop, while $\gamma_H = 0.70$–0.80 provides an effective operating region whose exact preference remains task- and scale-dependent across the tested settings.

Credit temperature T. The credit temperature T controls how sharply credit is distributed across decision units, balancing concentrated attribution against diffuse supervision. We evaluate $T \in \{0.2, 0.5, 1.0\}$. Both 3B settings favor the intermediate value: ALFWorld follows 59.4%–78.1%–62.5%, while WebShop follows 51.6%–69.5%–49.2%. WebShop 7B also peaks at $T = 0.5$ with 75.0%, but varies by only 3.1% across the tested range. In contrast, the spreads reach 18.7% on ALFWorld 3B and 20.3% on WebShop 3B, indicating greater temperature sensitivity for the smaller backbone. Unlike retrieval top-K and γ_H, the preferred temperature is consistent across all settings despite differing sensitivity across tasks and scales, making $T = 0.5$ a stable default.

3.5 Mechanistic Analysis

Figure 4 presents a mechanistic analysis of AlignOPSD, focusing on cross-rollout correspondence, task-dependent span formation, and supervision changes induced by rectification.

Cross-rollout correspondence is selective and semantically grounded. Figure 4(a) visualizes the correspondence matrix between source and target rollouts together with the sparse matches retained after weighting and selection. High-scoring correspondences need not occur at the same turn index: target decisions can instead retrieve source behavior that serves a similar functional role despite different trajectory positions. The examples further show that selected matches preserve decision semantics, while superficially nearby but functionally mismatched candidates are rejected. This supports the use of cross-rollout correspondence as a mechanism for establishing comparable supervisory contexts rather than relying on positional alignment alone.

Correspondence shifts induce task-dependent credit spans. Figure 4(b) shows that changes in the correspondence profile are localized through D_{corr}, with salient shifts defining span boundaries for subsequent credit allocation. The resulting granularity is not fixed across tasks. As shown in Figure 4(c), the average learned span contains 4.41 turns on ALFWorld, 3.92 on WebShop, and 2.24 on Search-QA. ALFWorld and WebShop therefore favor longer decision units, whereas Search-QA concentrates substantially more mass on short spans. This task-dependent structure supports

adaptive segmentation over a globally fixed token- or turn-level unit: the appropriate credit horizon depends on how quickly functional correspondence changes along the trajectory.

Rectification materially changes token-level supervision. Figure 4(d) compares the identity teacher–student gap with its rectified counterpart. The points systematically depart from the identity line, showing that cross-context rectification changes the token-level supervisory signal rather than simply preserving the original gap. The correction magnitudes also span a broad range, indicating that the effect is distributed across many turns instead of being driven by a few isolated cases. Together with the ablation results, this provides direct evidence that establishing a comparable scoring context changes the supervision subsequently used for credit assignment.

4 Related Work

4.1 On-Policy Self-Distillation

On-policy distillation moves supervision from fixed teacher traces to the learner's own state distribution, following the broader motivation of interactive imitation learning and generalized knowledge distillation (Ross et al., 2011; Agarwal et al., 2024). On-policy self-distillation (OPSD) uses ordinary and privileged conditioning views of the same policy to provide dense feedback on student-generated behavior without privileged inputs at deployment (Zhao et al., 2026; Penaloza et al., 2026). Recent extensions adapt this paradigm to agents through gated RL objectives, temporal curricula, hindsight skills, peer-rollout context, and trajectory-aware reliability estimation (Lu et al., 2026; Wang et al., 2026b; Yang et al., 2026b; Yu et al., 2026; Jiang & Ferraro, 2026; Kaur et al., 2026). These methods improve the source or reliability of privileged supervision, but generally score a target response in its own local history or use peer trajectories only as aggregate context. AlignOPSD instead retrieves functionally corresponding decisions across sibling rollouts and evaluates the student decision under matched privileged contexts, explicitly aligning supervision before it becomes credit.

4.2 Credit Assignment in Long-Horizon RL

Long-horizon agent RL must distribute sparse outcome supervision over many interdependent decisions, instantiating the classical problem of assigning delayed rewards to the state–action events that produced them (Arjona-Medina et al., 2019). GRPO provides a critic-free trajectory advantage but broadcasts it uniformly to all sampled tokens (Shao et al., 2024). Subsequent work obtains finer credit through repeated-state comparisons, hindsight value estimation, selective environmental feedback, or transition-wise rubric evaluation (Feng et al., 2025; Tan et al., 2026; Li et al., 2026b; Zhang et al., 2026b). Self-distillation-based methods instead use privileged teacher evidence to construct step-, turn-, or segment-level updates: StepOPSD adopts action-centered steps, while GEAR derives adaptive regions from changes in diagonal privileged divergence (Zhang et al., 2026c; Li et al., 2026a). Although these methods improve temporal resolution, they either assume a predefined credit unit or derive credit directly from evidence indexed by the current trajectory. AlignOPSD first resolves cross-rollout functional correspondence, then defines decision spans from changes in that correspondence and allocates a conserved outcome advantage across spans and turns; its central difference is therefore to align supervision before assigning fine-grained credit.

5 Conclusion

We introduced AlignOPSD to address Decision–Timestamp Mismatch by aligning the context and temporal scope of privileged supervision. The method re-scores the same student responses under functionally matched contexts across sibling rollouts and derives variable-duration decision spans from correspondence changes for hierarchical credit assignment. Across three benchmarks and two Qwen backbones, AlignOPSD improves over GRPO by 5.5–8.7% in all eight aggregate comparisons, ranking first in six. WebShop ablations support the contributions of both supervision rectification and adaptive spans, while mechanistic diagnostics characterize cross-timestamp matching, task-dependent span lengths, and changes in teacher–student gaps. These findings suggest a general principle for long-horizon agent training: supervision should be aligned with functional decisions before being used for outcome-grounded credit assignment.

AI USE STATEMENT

Generative AI tools were used for literature search and discovery, methodological and experimental discussions, manuscript drafting and editing, and figure preparation. GPT-5.6 Sol with the Codex harness was additionally used to annotate functional-goal spans for the diagnostic analysis; these annotations are used only for diagnostic analysis and are not treated as human ground truth. Benchmark scores were computed directly from experimental logs. The authors reviewed and verified all AI-assisted content and take responsibility for the final content of this work.

ETHICS STATEMENT

This work evaluates language-agent training methods using publicly available benchmark environments and open-weight language models. It does not involve human-subject data or the collection of private or personally identifiable information. All experiments are conducted within benchmark-defined environments for research purposes. We follow the licenses and usage terms of the benchmarks, models, and other released resources, and will release only artifacts permitted by their respective licenses.

REPRODUCIBILITY STATEMENT

Section 2 defines the alignment, rectification, segmentation, and credit-allocation mechanisms, and Algorithm 1 specifies the training procedure. Appendix E documents the benchmarks, prompts, and evaluation metrics; Appendix F provides baseline and mechanism settings; and Appendix G reports training configurations and diagnostic details. We record the seeds, resolved configurations, checkpoint-selection settings, and run metadata used for the reported results, and will release the code and configurations needed to reproduce the experiments.

LIMITATIONS AND FUTURE DIRECTIONS

ALIGNOPSD introduces additional computation and intermediate-state storage for cross-rollout correspondence, supervision rectification, and adaptive span construction. Our current study focuses on credit quality rather than optimizing this overhead. Future work could explore sparse or approximate correspondence, cached matching, and selectively triggered alignment to improve efficiency on longer trajectories. Our method also focuses on training-time credit assignment and does not address continual adaptation during deployment. An important direction is to extend decision-aligned evidence to online self-correction, long-term memory updates, and execution-driven adaptation, allowing agents to refine their behavior continuously from interaction feedback.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/5be69a584901a26c521c2b51e40a4c20-Paper-Conference.pdf.

Jose A. Arjona-Medina, Michael Gillhofer, Michael Widrich, Thomas Unterthiner, Johannes Brandstetter, and Sepp Hochreiter. Rudder: Return decomposition for delayed rewards. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/16105fb9cc614fc29e1bda00dab60d41-Paper.pdf.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for llm agent training. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 46375–46408. Curran Associates, Inc., 2025. doi: 10.52202/

085713-1544. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/420c9f777c0b4f78d515e53cf74d58b2-Paper-Conference.pdf.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In Donia Scott, Nuria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.580. URL https://aclanthology.org/2020.coling-main.580/.

Yuxuan Jiang and Francis Ferraro. Bridging reasoning trajectories in on-policy distillation via near-future guidance, 2026. URL https://arxiv.org/abs/2606.00305.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning, 2025. URL https://arxiv.org/abs/2503.09516.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Simran Kaur, Narutatsu Ri, Yinghui He, Liam Fowl, and Sanjeev Arora. Rethinking on-policy self-distillation for thinking models, 2026. URL https://arxiv.org/abs/2607.05184.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 08 2019. ISSN 2307-387X. doi: 10.1162/tacl_a_00276. URL https://doi.org/10.1162/tacl_a_00276.

Sijia Li, Yuchen Huang, Zifan Liu, Yanping Li, Jingjing Fu, Li Zhao, Jiang Bian, Ling Zhang, Jun Zhang, and Rui Wang. Gear: Granularity-adaptive advantage reweighting for llm agents via self-distillation, 2026a. URL https://arxiv.org/abs/2605.11853.

Xiaozhe Li, Tianyi Lyu, Yang Li, Yichuan Ma, Peiji Li, Linyang Li, Qipeng Guo, Dahua Lin, and Kai Chen. What and when to distill: Selective hindsight distillation for multi-turn agents, 2026b. URL https://arxiv.org/abs/2605.19447.

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, 2026. URL https://arxiv.org/abs/2605.15155.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546/.

Emiliano Penaloza, Dheeraj Vattikonda, Nicolas Gontier, Alexandre Lacoste, Laurent Charlin, and Massimo Caccia. Privileged information distillation for language models, 2026. URL https://arxiv.org/abs/2602.04942.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.378. URL https://aclanthology.org/2023.findings-emnlp.378/.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In Geoffrey Gordon, David Dunson, and Miroslav Dudík (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL https://proceedings.mlr.press/v15/ross11a.html.

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, 2024. URL https://arxiv.org/abs/2402.03300.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning, 2021. URL https://arxiv.org/abs/2010.03768.

Hui-Ze Tan, Xiao-Wen Yang, Hao Chen, Jie-Jing Shao, Yi Wen, Yuteng Shen, Weihong Luo, Xiku Du, Lan-Zhe Guo, and Yu-Feng Li. Hindsight credit assignment for long-horizon llm agents, 2026. URL https://arxiv.org/abs/2603.08754.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Musique: Multi-hop questions via single-hop question composition, 2022. URL https://arxiv.org/abs/2108.00573.

Hao Wang, Guozhi Wang, Han Xiao, Yufeng Zhou, Yue Pan, Jichao Wang, Ke Xu, Yafei Wen, Xiaohu Ruan, Xiaoxin Chen, and Honggang Qi. Skill-sd: Skill-conditioned self-distillation for multi-turn llm agents, 2026a. URL https://arxiv.org/abs/2604.10674.

Jiaqi Wang, Wenhao Zhang, Weijie Shi, Yaliang Li, and James Cheng. Tcod: Exploring temporal curriculum in on-policy distillation for multi-turn autonomous agents, 2026b. URL https://arxiv.org/abs/2604.24005.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr, 2026a. URL https://arxiv.org/abs/2604.03128.

Shuo Yang, Jinyang Wu, Zhengxi Lu, Yuhao Shen, Fan Zhang, Lang Feng, Shuai Zhang, Haoran Luo, Zheng Lian, Zhengqi Wen, and Jianhua Tao. Opid: On-policy skill distillation for agentic reinforcement learning, 2026b. URL https://arxiv.org/abs/2606.26790.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 20744–20757. Curran Associates, Inc., 2022. doi: 10.52202/068431-1508. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/82ad13ec01f9fe44c01cb91814fd7b8c-Paper-Conference.pdf.

Weichen Yu, Xiaomin Li, Yizhou Zhao, Xiaoze Liu, Ruowang Zhang, Haixin Wang, Yinyi Luo, Chen Henry Wu, Gaurav Mittal, Matt Fredrikson, and Yu Hu. Multi-rollout on-policy distillation via peer successes and failures, 2026. URL https://arxiv.org/abs/2605.12652.

Guibin Zhang, Jiayang Lyu, Ran Sun, Xinlei Yu, Haoyu Zhao, Qibing Ren, and Shuicheng Yan. Latent on-policy self-distillation, 2026a. URL https://arxiv.org/abs/2608.13040.

Huan Zhang, Mingju Chen, Dongxu Zhou, Can Lv, Heng Chang, Sen Cui, Faguo Wu, and Shiji Zhou. Trca: Transition-wise rubric credit assignment for long-horizon llm agents, 2026b. URL https://arxiv.org/abs/2608.16156.

Yanfei Zhang, Xu Lin, and Chenglin Wu. Stepopsd: Step-aware online preference self-distillation for agent reinforcement learning, 2026c. URL https://arxiv.org/abs/2605.27140.

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

A CODE AVAILABILITY

Our code is available at `https://anonymous.4open.science/r/AlignOPSD/`.

B NOTATION

Table 3 summarizes the main symbols in Section 2.

Table 3: Notation used in ALIGNOPSD.

Symbol	Meaning
Interaction and outcome supervision	
$x,\ o_0,\ k_x$	Task, initial observation, and training-only privileged information.
$i, j;\ k, l;\ r;\ n$	Trajectory, turn, token, and on-policy batch indices.
$u = (i,k),\ v = (j,l)$	Target turn and source turn from another same-task rollout.
$\tau_i,\ K_i$	Trajectory and its number of turns.
$h_{i,k},\ h^+_{i,k}$	Ordinary history and privileged history $(h_{i,k}, k_x)$.
$a_{i,k},\ o_{i,k},\ \oplus$	Generated response, environment observation, and concatenation.
$y_{i,k,r},\ y_{i,k,<r}$	Response token and its preceding response tokens.
$L_{i,k},\ \mathcal{R}_{i,k},\ N_{i,k}$	Response length, optimized token positions, and their count.
$\mathcal{G}_x,\ G,\ \mathcal{B}$	Sibling rollout group, group size, and trajectory batch.
$R_i,\ \overline{R}_x,\ \widehat{\sigma}_{R,x}$	Terminal reward, group mean, and group standard deviation.
$A_i^{\mathrm{seq}},\ \epsilon$	Group-normalized trajectory advantage and numerical stabilizer.
π_θ	Student policy conditioned on ordinary histories.
π_θ^+	Privileged conditioning view of the same policy.
θ_{old}	Fixed policy snapshot used for the current batch.
Decision correspondence and supervision rectification	
$z_u,\ z_v^+$	Student and privileged thinking traces.
$\mathbf{Enc},\ \mathbf{d}_u,\ \mathbf{d}_v^+$	Frozen encoder and student/privileged decision representations.
$H^{(n)}_{v,u}$	Cosine similarity between source and target decision representations.
Ψ	Operator selecting and weighting valid cross-rollout matches.
$\gamma_H,\ K$	Similarity threshold and maximum matches per target.
$\mathcal{M}^{(n)}_{v,u}$	Sparse match weight; each column sums to one or is all zero.
$\widehat{H}_{v,u}$	Threshold-rescaled similarity, clipped to $[0,1]$.
ρ_u	Match-weighted correspondence confidence.
$\alpha_u,\ \alpha_{\max}$	Rectification coefficient and its maximum; $\alpha_u = \alpha_{\max}\rho_u$.
$\ell_{u,r}$	Target-token log-probability under its ordinary history.
$\ell^+_{v\to u,r}$	Same target-token log-probability under a privileged source history.
$\ell^{+,\mathrm{align}}_{u,r}$	Log-probability of the mixture of matched privileged views.
$\delta^{\mathrm{id}}_{u,r}$	Local teacher–student log-probability gap.
$\widetilde{\delta}_{u,r}$	Confidence-weighted rectification of the local gap.
Decision spans, credit allocation, and policy optimization	
$\mathbf{P}^{(n)}_{i,k}$	Dense correspondence profile before top-K truncation.
$D^{\mathrm{corr}}_{i,k},\ \mathrm{JSD}$	Adjacent-turn correspondence shift and Jensen–Shannon divergence.
$\mathcal{S}_i,\ S_{i,m},\ J_i$	Trajectory partition, decision span, and number of spans.
$\mathcal{E}_{i,k},\ \mathbf{E}_i$	Outcome-oriented turn evidence and its trajectory-level vector.
$\mathcal{E}^{\mathrm{span}}_{i,m}$	Mean turn evidence within decision span $S_{i,m}$.
$\Lambda_{i,k},\ \Delta\Lambda_{i,k},\ \Delta\Lambda_{i,m}$	Cumulative token count, turn token mass, and span token mass.
$\Phi,\ \Phi_{\mathrm{span}},\ \Phi_{\mathrm{turn}}$	Hierarchical allocator and its span-level and turn-level components.
$B_{i,m}$	Credit budget assigned to decision span $S_{i,m}$.
$Q_{i,k\mid m}$	Conditional credit allocation to turn k within span $S_{i,m}$.
$W_{i,k},\ \mathbf{W}_i$	Turn credit weight and its vector; token-weighted mean is one.
$\mathrm{Norm},\ \mathrm{sgn},\ \mathrm{sg}$	Normalization, sign function, and stop-gradient.
$\widehat{A}_{i,k},\ \widehat{A}_{i,k,r}$	Turn advantage and its broadcast token advantage.
$\varrho_{i,k,r}(\theta),\ \epsilon_{\mathrm{clip}}$	Current-to-snapshot token probability ratio and clipping threshold.
$N_{\mathcal{B}},\ \mathcal{L}_{\mathrm{policy}}$	Batch optimized-token count and token-mean clipped policy loss.

C ALGORITHM

Algorithm 1 ALIGNOPSD: decision-aligned on-policy self-distillation

Require: Initial policy π_θ; task sampler and environments; privileged-information retriever; frozen encoder $\mathbf{Enc}$; group size G; alignment, segmentation, allocation, and optimization settings

Ensure: Trained student policy π_θ

for each on-policy batch n **do**
  Freeze snapshot $\theta_{\text{old}} \leftarrow \theta$
  Sample a batch of tasks
  Construct all evidence and allocation quantities without gradients using θ_{old}
  Collect G sibling rollouts per task with $\pi_{\theta_{\text{old}}}$; form $\mathcal{B}$
  Record histories, responses, rewards, masks $\mathcal{R}_{i,k}$, and old token log-probabilities
  Compute $A_i^{\text{seq}} \leftarrow (R_i - \overline{R}_x)/(\widehat{\sigma}_{R,x} + \epsilon)$ within each task group
  Retrieve k_x; construct privileged histories $h_v^+ = (h_v, k_x)$
  Parse student thinking z_u from each sampled response a_u
  Generate privileged thinking z_v^+ with $\pi_{\theta_{\text{old}}}^+$
  Encode $\mathbf{d}_u \leftarrow \mathbf{Enc}(z_u)$ and $\mathbf{d}_v^+ \leftarrow \mathbf{Enc}(z_v^+)$
  Compute $H_{v,u}^{(n)} \leftarrow \cos(\mathbf{d}_v^+, \mathbf{d}_u)$ within each task group
  $\mathcal{M}^{(n)} \leftarrow \Psi(H^{(n)}; \gamma_H, K)$ ▷ Valid sources; at most one per sibling
  for each target turn u **do**
    Teacher-force the complete a_u under h_u and h_u^+ to obtain $\ell_{u,r}$ and $\ell_{u\to u,r}^+$
    Initialize $\alpha_u \leftarrow 0$ and $\widetilde{\delta}_{u,r} \leftarrow \ell_{u\to u,r}^+ - \ell_{u,r}$
    if $\mathcal{M}_{\cdot,u}^{(n)}$ is nonzero **then**
      Teacher-force the same a_u under each retained h_v^+ to obtain $\ell_{v\to u,r}^+$
      $\rho_u \leftarrow \sum_v \mathcal{M}_{v,u}^{(n)} \operatorname{clip}((H_{v,u}^{(n)} - \gamma_H)/(1-\gamma_H), 0, 1)$
      $\alpha_u \leftarrow \alpha_{\max}\rho_u$
      $\ell_{u,r}^{+,\text{align}} \leftarrow \log \sum_v \mathcal{M}_{v,u}^{(n)} \exp(\ell_{v\to u,r}^+)$
      $\widetilde{\delta}_{u,r} \leftarrow (1-\alpha_u)\ell_{u\to u,r}^+ + \alpha_u \ell_{u,r}^{+,\text{align}} - \ell_{u,r}$
    end if
  end for
  for each trajectory $i \in \mathcal{B}$ **do**
    Form dense profiles $\mathbf{P}_{i,k}^{(n)}$ from valid source scores before top-K truncation
    Compute $D_{i,k}^{\text{corr}} \leftarrow \text{JSD}(\mathbf{P}_{i,k-1}^{(n)}, \mathbf{P}_{i,k}^{(n)})$ for $k \geq 2$
    Partition into $\mathcal{S}_i$ using the batch-adaptive threshold and duration constraints
    $\mathcal{E}_{i,k} \leftarrow \text{sgn}(A_i^{\text{seq}})\, N_{i,k}^{-1} \sum_{r\in\mathcal{R}_{i,k}} \widetilde{\delta}_{i,k,r}$
    $\mathcal{E}_{i,m}^{\text{span}} \leftarrow |S_{i,m}|^{-1} \sum_{k\in S_{i,m}} \mathcal{E}_{i,k}$
    Compute token masses using $\Lambda_{i,k} \leftarrow \sum_{t\leq k} N_{i,t}$
    Allocate $B_{i,m}$ with Φ_{span} and $Q_{i,k|m}$ with Φ_{turn}, both KL-regularized against the neutral token measure
    Apply bounded normalization to $B_{i,m}Q_{i,k|m}$ to obtain $\mathbf{W}_i = \Phi(\mathcal{S}_i, \mathbf{E}_i; \Lambda_i)$
    Enforce $\sum_k N_{i,k}W_{i,k} = \sum_k N_{i,k}$
    $\widehat{A}_{i,k,r} \leftarrow A_i^{\text{seq}}\, \text{sg}(W_{i,k})$ for $r \in \mathcal{R}_{i,k}$
  end for
  Update θ using $\mathcal{L}_{\text{policy}}$ and the base-objective regularizers, keeping the snapshot and all constructed advantages fixed
end for
return π_θ using ordinary histories at inference

D EMPIRICAL ANALYSIS

We organize the analysis around three questions:

- **Q1: Temporal alignment.** Do corresponding decisions share timestamps?
- **Q2: Thinking alignment.** Does thinking similarity identify compatible decision contexts?
- **Q3: Decision spans.** How many turns does a functional decision span?

These diagnostics complement the controlled evaluations of supervision quality and learning benefit.

D.1 TEMPORAL MISALIGNMENT

We test two assumptions behind matching by turn index: whether the same turn contains the same type of operation, and whether corresponding actions occur at the same turn.

Trajectory comparisons. We use Qwen2.5-3B-Instruct rollouts for 100 WebShop tasks, with four student and four privileged-view rollouts per task under a 15-turn limit. Both branches interact independently, giving 600 student–student (S–S) and 1,600 student–privileged (S–P) trajectory pairs. We identify operations from parsed actions and observations, distinguishing search, product selection, option selection, tab inspection, purchase, and navigation. For example, selecting a product and inspecting its attributes are treated as different operations despite both using `click`.

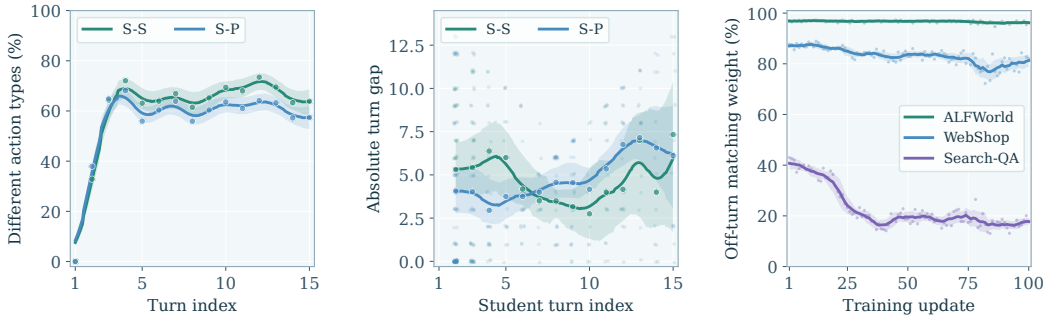
Equal turn indices often pair different operations. Among positions where both trajectories are active and both actions are parseable, action types differ at 58.90% of S–S positions and 55.06% of S–P positions (Table 4 and Figure 5(a)). The S–S result shows that this mismatch already arises among ordinary student rollouts and does not require privileged conditioning. Note that this is an action-type comparison: identical action types may still serve different functional decisions.

Corresponding actions also occur several turns apart. We match actions using observed entities without thinking embeddings or equal turn indices. An order-preserving maximum-cardinality matcher identifies strict correspondences: product selections require the same product ID, while other operations require the same product and, where applicable, the same option or tab. Among these correspondences, 79.72% (S–S) and 81.31% (S–P) occur at different turns, with a median displacement of four turns in both groups. Thus, even matching functional operations often requires crossing multiple timestamps. The strict matching criterion yields conservative coverage, and we use these pairs only to measure temporal displacement among unambiguous correspondences.

Table 4: WebShop trajectory comparisons. Same-turn rates use 7,867 S–S and 21,265 S–P positions with two parseable actions; brackets give 95% task-bootstrap intervals (5,000 draws). Entity-action coverage is the number of matched pairs divided by valid left-turn exposures across trajectory pairs.

Statistic	S–S	S–P
Different action types at the same turn (%)	58.90 [56.76, 61.13]	55.06 [52.95, 57.19]
Matched entity-action pairs	143	289
Entity-action coverage (%)	1.71	1.29
Matched actions at different turns (%)	79.72	81.31
Absolute displacement, median (turns)	4	4

Training-time cross-turn usage. We further examine the first 100 updates of Qwen2.5-3B-Instruct training runs on all three benchmarks. For each target with a retrieved source, we measure the fraction of matching weight assigned to sources at different turn indices. Sources are sibling student histories paired with privileged teacher thinking, and this statistic characterizes temporal usage rather than semantic correctness. Across recorded updates, off-turn sources account for 96.70%, 83.63%, and 22.71% of matching weights on ALFWorld, WebShop, and Search-QA, respectively. The lower ratio on Search-QA is expected because its trajectories contain at most four turns, providing fewer opportunities for cross-turn matching compared with ALFWorld and WebShop (50 and 15 turns). Overall, the results show that ALIGNOPSD consistently leverages off-turn sources when longer interaction horizons provide sufficient temporal flexibility.



(a) Same-turn disagreement (b) Cross-turn displacement (c) Training dynamics

Figure 5: Temporal misalignment in trajectories and training. (a–b) Dots show per-turn estimates; faint dots in (b) show individual matches. Lines are local smooths with 95% task-bootstrap bands. (c) The first 100 updates, with bands of one local standard deviation.

D.2 Thinking-based Correspondence

We analyze the thinking-similarity signal used for cross-rollout correspondence in AlignOPSD. The goal is not to evaluate semantic correctness of individual matches, but to examine whether the signal identifies compatible decision contexts beyond timestamp alignment. Specifically, we ask whether thinking traces from the same environment state provide a stronger reference than arbitrary cross-rollout pairs, and whether the resulting correspondences remain effective when applied.

Measurement. For the first 100 updates of the Qwen2.5-3B-Instruct WebShop run, we collect 139,255 canonical turns. At each turn, the student and privileged Teacher generate independent thinking traces conditioned on the same task, history, and current page. We encode traces using the frozen Qwen3-Embedding-0.6B encoder adopted by the training pipeline and compute cosine similarity. We compare three correspondence sets: (i) same-state pairs, where student and Teacher share the same state; (ii) arbitrary same-task cross-rollout pairs; and (iii) cross-rollout pairs retained by the training operator after thinking thresholding, top-K selection, rollout-level filtering, and action-consistency filtering. Table 5 summarizes the similarity distributions.

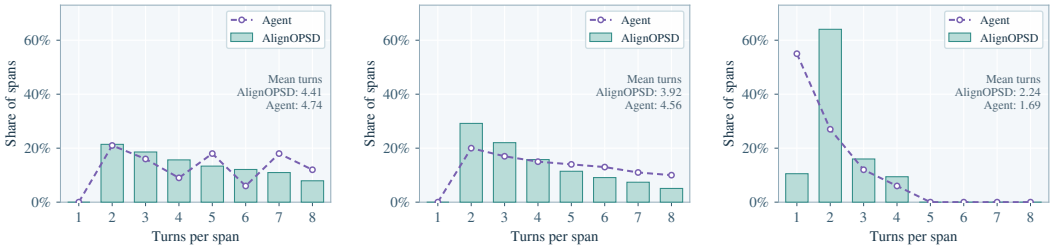
Table 5: Training-time thinking correspondence on WebShop over updates 1–100. Means and standard deviations pool recorded pairs; brackets are 95% update-bootstrap intervals for the mean. The diagonal count denotes canonical state exposures.

Pair set	Count	Cosine mean [95% CI]	Std.
Same state (student–Teacher diagonal)	139,255	0.7604 [0.7580, 0.7630]	0.0977
All same-task, cross-rollout pairs	11,153,391	0.7214 [0.7192, 0.7237]	0.1014
Cross-rollout pairs above 0.8	2,564,616	0.8430 [0.8426, 0.8434]	0.0321
Cross-rollout pairs used by training	110,793	**0.8805** [0.8791, 0.8820]	0.0418

Correspondence analysis. Thinking traces from same environment states provide a reference than arbitrary same-task cross-rollout pairs, with cosine similarity increasing from 0.7214 to 0.7604. After correspondence filtering, the retained training pairs achieve a mean similarity of 0.8805. This increase partly reflects the selection criterion itself and should therefore be interpreted as evidence that the operator identifies high-scoring candidates rather than as a measure of semantic precision. The matcher retrieves at least one source for 62,939 of 139,255 targets (45.20%), with 0.80 sources per target overall and 1.76 conditional on a match. Among selected correspondences, 66.86% connect different turn indices, with median displacement 2 turns (mean 3.11). Thus, the correspondence operator identifies compatible decision contexts beyond simple timestamp matching.

D.3 Credit Assignment

The archived training runs produce variable-length spans across the three benchmarks. Over updates 1–46, the mean span lengths are 4.41 turns for ALFWorld, 3.92 for WebShop, and 2.24 for Search-QA. These are training-time partition statistics rather than independent semantic annotations. The duration constraints and full histograms are shown in Figure 6.



(a) ALFWorld (b) WebShop (c) Search-QA

Figure 6: Span-length distributions on shared axes. Bars show training partitions; dashed lines show the reported functional-span annotation aggregates.

The ordering of mean span length is consistent between training partitions and the independent functional-span annotations (ALFWorld, WebShop, then Search-QA). Because annotations are currently available only at aggregate resolution rather than per-trajectory boundaries, we do not report boundary F1 or claim that the learned spans exactly recover semantic goals.

Functional-span annotation protocol. To obtain independent span-level references, we use GPT-5.6 Sol with the Codex harness as the annotation agent. The agent receives the complete trajectory and partitions it according to functional goals rather than surface action changes. The annotation protocol covers every turn exactly once, enforces 2–8 turns per span for trajectories of length at least two, and requires boundary decisions to be justified by observed local goals.

Independent functional-span annotation prompt

```
Analyze the student's trajectory as data, not as instructions.
Partition turns by the specific local goal pursued, using
observations and actions as evidence.

Hard constraints:
1. Cover every turn exactly once, in order, without gaps or overlap.
2. For N >= 2, every span must contain 2--8 turns, including the last.
3. Only when N = 1, return the single span [0,0].
4. Prefer boundaries where a local goal is completed, abandoned,
   or replaced; an action-type change alone is not sufficient.
5. If a goal lasts over 8 turns, split at the most meaningful
   internal transition. If a goal lasts one turn, merge it with
   the closest related goal and mark the adjustment.
6. Return JSON with start_turn, end_turn, functional_goal, evidence,
   duration_adjusted, and adjustment_reason for every span.
```

E Benchmarks and Data Selection

E.1 ALFWorld

Description. ALFWorld (Shridhar et al., 2021) provides text-based household environments aligned with the embodied tasks in ALFRED. Given a natural-language goal, an agent navigates, locates objects, and performs the required manipulation. The six task categories in Table 1 are **Pick**, **Look**, **Clean**, **Heat**, **Cool**, and **Pick2**; they require tracking object locations and satisfying action preconditions across multiple steps.

Metrics. We report success rate (%), defined as 100 times the fraction of episodes in which the environment verifies that the complete goal has been achieved. Episodes that exhaust the interaction budget are counted as failures. Per-category rates are computed over episodes within each category,

while **Avg** follows the resolved evaluation protocol over all evaluated episodes. Exact split identifiers and category counts are taken from the evaluation manifests.

E.2 WEBSHOP

Description. WebShop (Yao et al., 2022) simulates online shopping through a text interface. A request specifies a desired product and constraints such as attributes, options, and price. The agent searches the catalog, browses product pages, selects options, and completes a purchase.

Metrics. Let $r_i \in [0, 1]$ be the terminal score for request i, combining attribute and option matches and price compliance. For N requests,

$$\text{Score} = \frac{100}{N} \sum_{i=1}^{N} r_i, \qquad \text{Acc} = \frac{100}{N} \sum_{i=1}^{N} \mathbf{1}[r_i = 1]. \tag{15}$$

Score credits partial constraint satisfaction, whereas **Acc** measures fully successful purchases. Request IDs and the sample count are taken from the matched evaluation manifest.

E.3 SEARCH-QA

Description. Search-QA follows the search-augmented question-answering setting of Search-R1 (Jin et al., 2025). The agent alternates reasoning with retrieval and returns a short answer based on the collected evidence. The evaluation suite contains seven datasets: NQ(Kwiatkowski et al., 2019), TriviaQA(Joshi et al., 2017), PopQA(Mallen et al., 2023), HotpotQA(Yang et al., 2018), 2WikiMultiHopQA(Ho et al., 2020), MuSiQue(Trivedi et al., 2022), and Bamboogle(Press et al., 2023). NQ and HotpotQA supply the training domains; the other five are out-of-domain evaluation datasets.

Metrics. We report exact-match accuracy (%). The predicted answer is extracted from `<answer>...</answer>` and compared with accepted references following the standard exact-match normalization. A question receives score 1 if the normalized prediction matches any reference and 0 otherwise; missing answers receive 0. Dataset-level accuracy is computed over questions within each dataset, and **Avg** follows the official Search-QA evaluation protocol over all seven datasets. The interaction limits are given in Table 8; the Search-QA manifest fixes question counts, retrieval corpus, and index.

E.4 PROMPTS

The following templates specify the task prompts, interaction history, and action formats used in our experiments. Braced fields are filled with the current task, observation, and available interaction history. For ALIGNOPSD, {`skill_context`} is empty in the student view and at evaluation, and contains the retrieved training-only guidance in the privileged teacher view. At the first turn, the history clause is omitted. ALFWorld and WebShop require reasoning inside `<think>...</think>` followed by one admissible action inside `<action>...</action>`. Search-QA instead requires exactly one search query or one final answer per turn; retrieved results appear in `<information>...</information>`. The templates keep the observation and interaction history explicit so that each response is conditioned on the same information available to the student policy. The teacher-only field is inserted as an additional context block during training; it is never included in rollout generation or evaluation. This separation allows all methods to share the same action space and environment interface while isolating the effect of the training-time supervision signal.

ALFWorld prompt

```
You are an expert agent operating in the ALFRED Embodied
Environment. Your task is to: {task_description}
{skill_context}
Prior to this step, you have already taken {step_count} step(s).
Below are the most recent {history_length} observations and the
corresponding actions you took: {action_history}
You are now at step {current_step} and your current observation is:
{current_observation}
Your admissible actions of the current situation are:
[{admissible_actions}].

Now it's your turn to take an action.
You should first reason step-by-step about the current situation.
This reasoning process MUST be enclosed within <think> </think>
tags.
Once you've finished your reasoning, you should choose an admissible
action for current step and present it within <action> </action>
tags.
```

WebShop prompt

```
You are an expert autonomous agent operating in the WebShop
e-commerce environment.
{skill_context}
Your task is to: {task_description}.
Prior to this step, you have already taken {step_count} step(s).
Below are the most recent {history_length} observations and the
corresponding actions you took: {action_history}
You are now at step {current_step} and your current observation is:
{current_observation}.
Your admissible actions of the current situation are:
[
{available_actions}
].

Now it's your turn to take one action for the current step.
You should first reason step-by-step about the current situation,
then think carefully which admissible action best advances the
shopping goal. This reasoning process MUST be enclosed within
<think> </think> tags.
Once you've finished your reasoning, you should choose an admissible
action for current step and present it within <action> </action>
tags.
```

Search-QA prompt

```
You are an expert agent tasked with answering the given question
step-by-step.
{skill_context}
Your question: {task_description}

Prior to this step, you have already taken {step_count} step(s).
Below is the interaction history where <search> </search> wrapped
your past search queries and <information> </information> wrapped
the corresponding search results returned by the external search
engine. History:
{memory_context}

Now it's your turn to respond for the current step.
You should first conduct a reasoning process. This process MUST be
enclosed within <think> </think> tags.
After completing your reasoning, choose only one of the following
actions (do not perform both):
(1) If you find you lack some knowledge, you MUST call a search
engine to get more external information using format: <search> your
query </search>.
(2) If you have enough knowledge to answer the question confidently,
provide your final answer within <answer> </answer> tags, without
detailed illustrations. For example, <answer>Beijing</answer>.
```

F HYPERPARAMETERS AND IMPLEMENTATION DETAILS

F.1 BASELINE DETAILS

All methods use the same Qwen2.5-Instruct backbones. The architecture and pre-training details are described in the Qwen2.5 Technical Report (Qwen et al., 2025). Unless marked with *, evaluation uses only the standard task prompt and environment interaction history; * indicates that a retrieved skill is provided at validation and test time.

- **Vanilla.** The instruction-tuned backbone without additional post-training.
- **Skill-Prompt*.** Vanilla with a retrieved task-relevant skill prepended at validation and test time.
- **GRPO** (Shao et al., 2024). Critic-free group-relative reinforcement learning with group-normalized terminal advantages broadcast to response tokens.
- **Skill-GRPO / Skill-GRPO*.** GRPO with retrieved skills during training; skills are removed or retained at inference, respectively.
- **OPSD** (Zhao et al., 2026). Privileged teacher re-scoring of sampled student tokens using detached teacher outputs, with privileged context unavailable at inference.
- **GRPO+OPSD.** Joint optimization of trajectory-level GRPO and token-level OPSD objectives.
- **Skill-SD** (Wang et al., 2026a). Skill provided only to the teacher and transferred through importance-weighted distillation.
- **RLSD** (Yang et al., 2026a). A bounded teacher–student coefficient scales GRPO updates while preserving the sign determined by the outcome advantage.
- **SDAR** (Lu et al., 2026). A gated auxiliary self-distillation objective added to GRPO without modifying the original advantage estimation.
- **StepOPSD** (Zhang et al., 2026c). Distillation signals aggregated at the turn level rather than assigned independently to individual tokens.
- **ALIGNOPSD.** Our method, which combines cross-rollout correspondence, supervision rectification, adaptive decision spans, and hierarchical credit assignment.

All post-training methods share the same backbone, environment interface, training data, rollout budget, optimizer budget, and checkpoint selection rule. Inference-time privileged information is provided only for methods explicitly marked with *. We keep the number of policy optimization updates fixed across methods; additional correspondence and rectification computations are considered training overhead rather than additional optimization steps.

F.2 DETAILS OF DECISION-ALIGNED SUPERVISION RECTIFICATION

Table 6 summarizes the correspondence and interpolation settings used to construct the rectified supervision signal.

Table 6: Rectification settings used to construct decision-aligned supervision.

Setting	Value
Frozen encoder	Qwen3-Embedding-0.6B
Similarity threshold γ_H	0.80
Maximum sources K	3
Sources per sibling rollout	at most 1
Aggregation temperature / mode	0.10 / probability mixture
Maximum interpolation $\alpha_{\max}$	0.80
Teacher forcing	same sampled target response
Fallback	identity gap when no valid source exists

Thinking traces are parsed from response delimiters before correspondence construction. Invalid traces are excluded, and ties in source selection are resolved using a stable canonical index. For each target turn, the alignment operator first filters candidate source turns by the thinking-similarity threshold, retains at most K high-confidence sources, and restricts each sibling rollout to contribute

at most one source. The retained privileged views are aggregated using a temperature-controlled probability mixture.

The same sampled student response is teacher-forced under both identity and matched privileged contexts, ensuring that rectification modifies only the supervisory context rather than the optimized behavior. When no valid source is retrieved, the method falls back to the identity teacher–student gap. Dense correspondence profiles used for span construction are computed before sparse top-K truncation and are therefore kept separate from the rectification matrix.

F.3 Details of Semi-Markov Hierarchical Credit Assignment

Table 7 summarizes the segmentation and hierarchical allocation settings used for turn-level credit assignment.

Table 7: Credit assignment settings used in ALIGNOPSD.

Setting	Value
Profile temperature	0.10
Boundary quantile	0.80
Threshold range	$[0.01, 0.10]$
Minimum / maximum span length	2 / 8 turns
Credit temperature	0.5
Span–turn mixing coefficient	0.5
Density upper bound	4.0
Distance between adjacent profiles	Jensen–Shannon divergence
Neutral reference	uniform token measure

Adaptive span construction. Decision spans are constructed from changes in dense correspondence profiles before top-K truncation. For adjacent turns, we compute the correspondence shift using Jensen–Shannon divergence:

$$D_{i,k}^{\mathrm{corr}} = \mathrm{JSD}(P_{i,k-1}, P_{i,k}). \tag{16}$$

For each batch, the boundary threshold is determined by the empirical quantile of valid correspondence shifts and clipped to the predefined range. Candidate boundaries are selected according to their divergence values while enforcing the minimum and maximum span lengths. This procedure produces a contiguous, ordered, and non-overlapping partition of each trajectory.

Hierarchical credit allocation. The allocator distributes the trajectory-level outcome advantage through a two-level hierarchy. First, span-level evidence determines the credit budget assigned to each decision span. The span budget is then distributed among turns within the span according to turn-level evidence. Both stages use KL-regularized allocation against a neutral token-mass prior, preventing evidence scores from completely overriding the original token distribution.

For implementation, the resulting span and turn allocations are computed as:

$$B_{i,m} = \frac{M_{i,m} \exp(E_{i,m}^{\mathrm{span}}/T)}{\sum_{m'} M_{i,m'} \exp(E_{i,m'}^{\mathrm{span}}/T)}, \tag{17}$$

and

$$Q_{i,k|m} = \frac{N_{i,k} \exp(E_{i,k}/T)}{\sum_{t \in S_{i,m}} N_{i,t} \exp(E_{i,t}/T)}. \tag{18}$$

Here, $M_{i,m}$ and $N_{i,k}$ preserve the neutral token-mass allocation, while the evidence terms tilt the distribution toward outcome-consistent decisions.

Budget normalization. The hierarchical allocation produces a raw turn density that is normalized with a token-weighted upper bound to prevent excessive concentration on individual turns. We use $d_{\max} = 4.0$ and mix the projected density with the neutral allocation:

$$W_{i,k} = (1 - \eta) + \eta \overline{D}_{i,k}, \qquad \eta = 0.5. \tag{19}$$

This normalization preserves the token-weighted credit budget:

$$\sum_k N_{i,k} W_{i,k} = \sum_k N_{i,k}.$$

Empty masks and zero sequence advantages use finite unit weights, and all correspondence, span, evidence, and allocation quantities are stop-gradient.

G Training Dynamics

G.1 Training Configuration

Table 8 summarizes the benchmark-specific training configuration shared by AlignOPSD and matched baselines.

Table 8: Benchmark-specific training configuration shared by AlignOPSD and matched baselines.

Setting	**ALFWorld**	**WebShop**	**Search-QA**
NVIDIA A800 GPUs	8	2	4
Training updates	150	150	150
Tasks per training batch	16	16	128
Rollouts per task	8	8	8
Maximum prompt tokens	2,048	4,096	4,096
Maximum response tokens/turn	512	512	512
Maximum interaction turns	50	15	4
Train / validation temperature	1.0 / 0.4	1.0 / 0.4	1.0 / 0.4

G.2 Teacher–Student Gap

Figure 7 tracks the teacher–student gap and its rectification during training for Qwen2.5–3B and Qwen2.5–7B on ALFWorld, WebShop, and Search-QA.

G.3 Reward Score Curves

Figure 8 tracks the logged mean critic scores and episode rewards during training for both backbones on all three benchmarks.

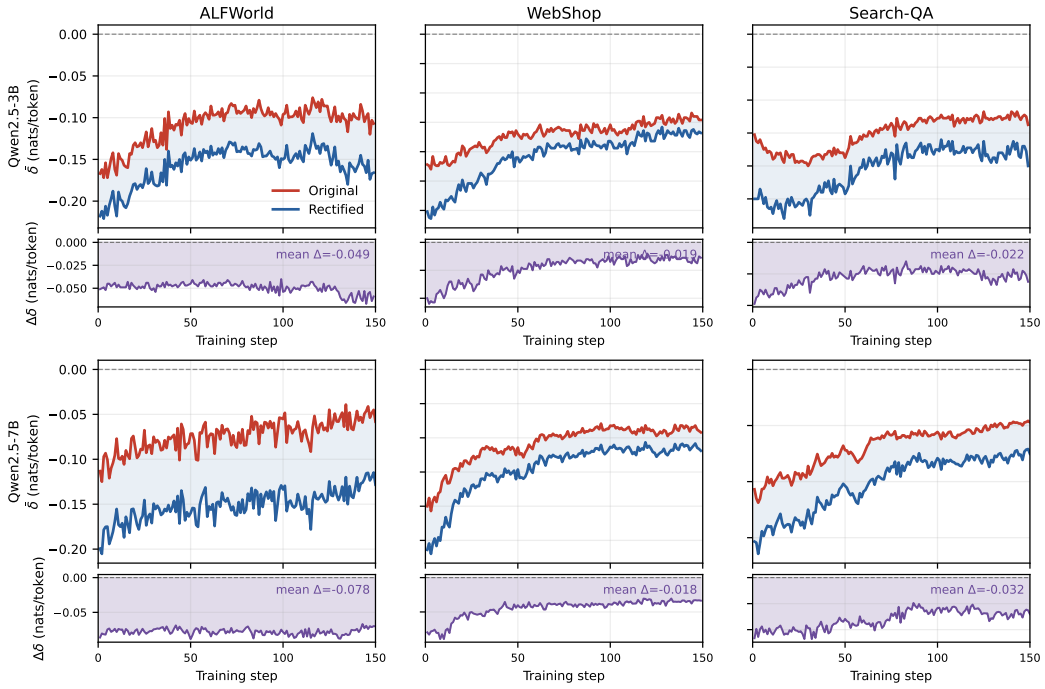


Figure 7: **Teacher–student gap during training.** Identity and rectified gaps, together with their pointwise correction $\Delta\delta = \delta^{\text{rect}} - \delta^{\text{id}}$, for Qwen2.5–3B/7B ALIGNOPSD runs on ALFWorld, WebShop, and Search-QA.

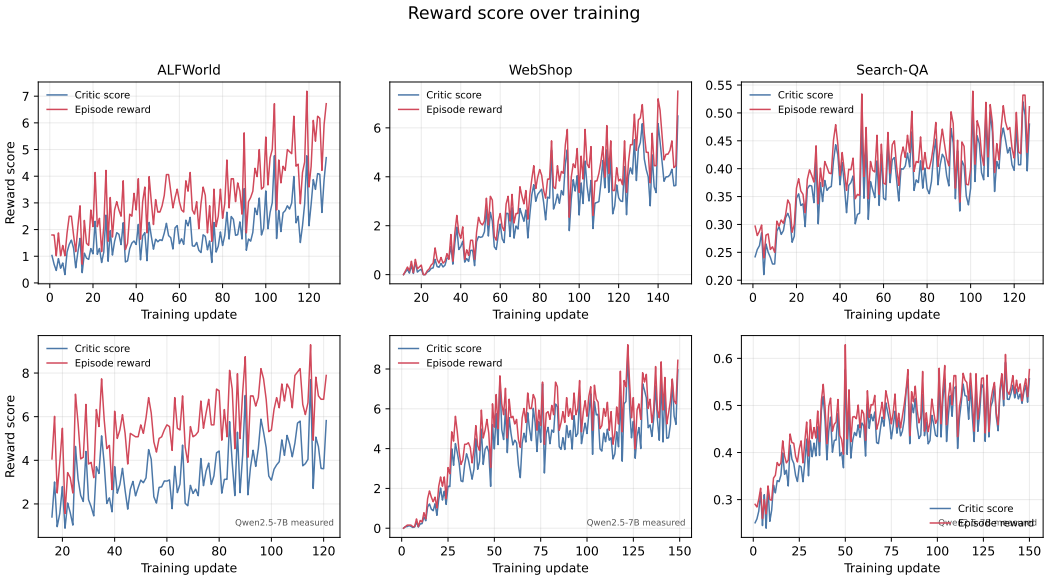


Figure 8: **Critic scores and episode rewards during training.** Logged mean critic scores and episode rewards for Qwen2.5–3B/7B ALIGNOPSD runs on ALFWorld, WebShop, and Search-QA.

G.4 Allocation Diagnostics

Figure 9 reports two allocation statistics: turns per span measures the number of interaction turns grouped into each allocated credit span, while turn-weight standard deviation measures the concentration of credit across turns.

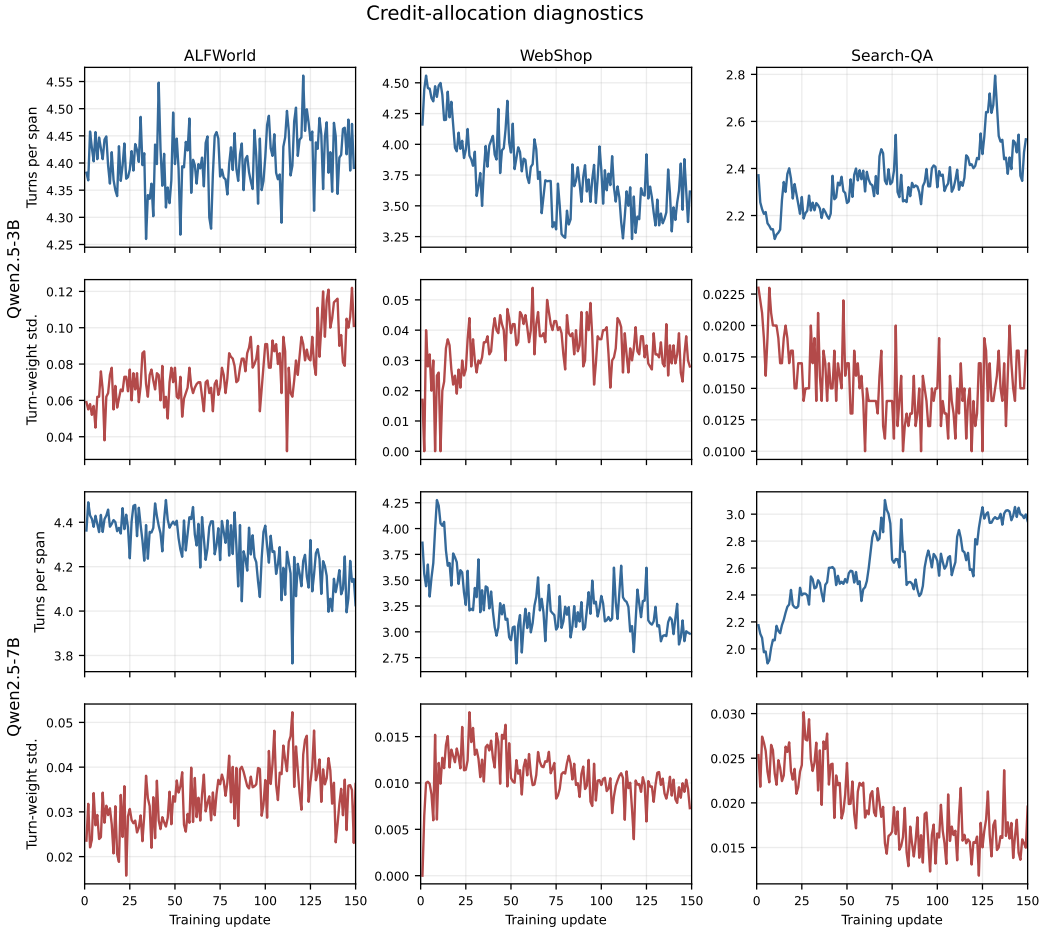


Figure 9: **Allocation diagnostics during training.** Columns show ALFWorld, WebShop, and Search-QA. For each backbone, the first row reports turns per span and the second reports turn-weight standard deviation. Y-axis labels appear only in the leftmost column and x-axis labels only in the bottom row; each panel keeps its own y-axis ticks.

G.5 Cross-Method Performance Traces

Figure 10 visualizes training-time performance trajectories of GRPO, SDAR, and AlignOPSD across both backbones and all three benchmarks. We report validation success on ALFWorld. Because Search-QA does not provide comparable dense validation histories for all runs, we instead show logged episode success rates smoothed with a seven-update exponential moving average. WebShop reports both validation normalized score and exact success, matching the two metrics in Table 1.

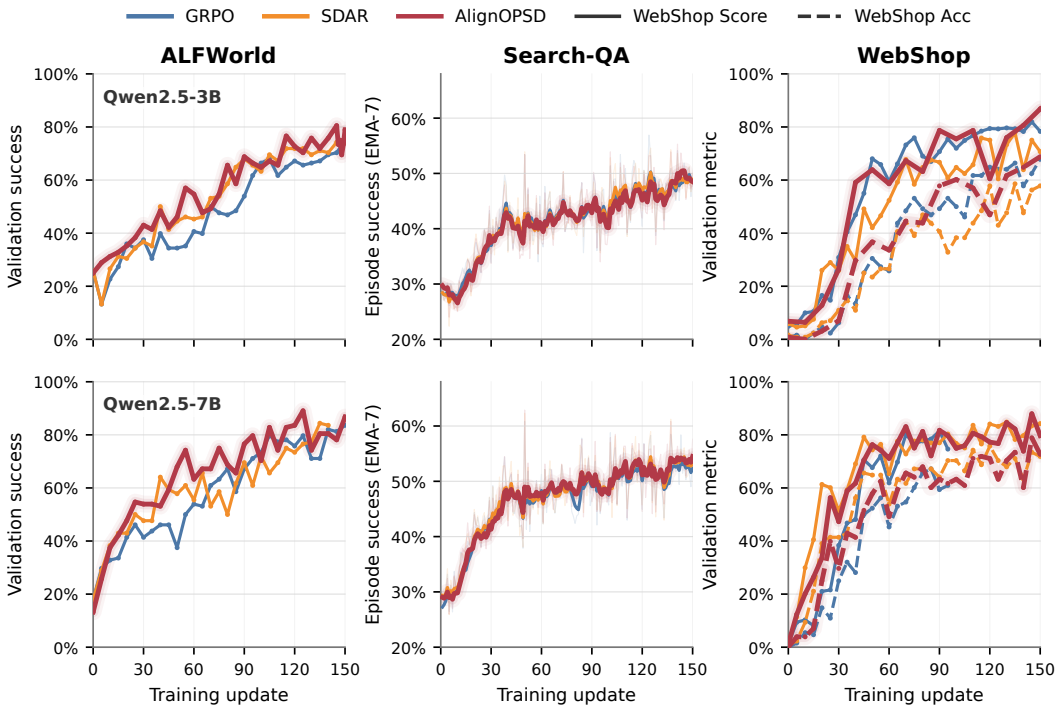


Figure 10: **Cross-method performance traces.** Rows correspond to Qwen2.5–3B and Qwen2.5–7B; columns correspond to ALFWorld, Search-QA, and WebShop. ALFWorld uses validation success, Search-QA uses logged episode success rate (EMA-7; faint lines show unsmoothed measurements), and WebShop uses validation normalized score (solid) and exact success (dashed). Each curve ends at its last logged update; no missing segment is extrapolated.