ECOD: EXPERIENCE-CALIBRATED ON-POLICY DISTILLATION FOR MULTI-TURN AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse and delayed rewards make reinforcement learning challenging for long-horizon language agents. On-policy distillation (OPD) complements environmental rewards with dense teacher feedback on student-generated contexts. Standard OPD can misalign supervision with relative teacher competence, student trajectory outcomes, and the temporal structure of task execution. Recent variants adjust guidance through divergence-based weighting and temporal schedules, while joint calibration from task-completion experience remains underexplored. We introduce Experience-Calibrated On-Policy Distillation (ECOD), which allocates teacher supervision at the task, trajectory, and turn levels. ECOD compares teacher and student success-rate posteriors under matched task conditions, restricts direct supervision to failed student trajectories, and derives a temporal prior from successful teacher lengths. The resulting weight applies to teacher losses or teacher-derived advantages. We evaluate ECOD on ALFWorld and WebShop with Qwen2.5-1.5B-Instruct and Qwen2.5-7B-Instruct. Under matched teacher settings, ECOD improves standard OPD by up to 4.06% in relative ALFWorld unseen success. Across both model scales, ECOD-assisted SOD, TCOD-F2B, and ATOD configurations yield average paired relative gains of 6.20% on ALFWorld unseen and 7.05% on WebShop, with WebShop gains ranging from 2.29% to 21.60%. These results support experience-calibrated supervision allocation as a reusable mechanism across distinct distillation objectives.

1 INTRODUCTION

Long-horizon language agents must coordinate sequential decisions under partial observability, often with sparse or delayed environmental rewards (Yao et al., 2022b). This makes credit assignment across extended interaction trajectories a central challenge for reinforcement learning (RL). Methods such as GiGPO improve multi-turn policy optimization, but sparse rewards still provide limited feedback on individual decisions (Feng et al., 2025).

On-policy distillation (OPD) complements sparse environmental rewards with dense supervision from a fixed teacher evaluated on student-generated contexts (Agarwal et al., 2024). Standard OPD can suffer from teacher supervision misalignment when guidance overlooks the teacher's relative competence, student trajectory outcomes, or the temporal structure of task execution. Recent variants make teacher guidance more adaptive through divergence-based weighting, temporal curricula, and turn-aware scheduling (Zhong et al., 2026; Wang et al., 2026; Tan et al., 2026). These methods improve specific aspects of supervision, but their adaptation remains tied to individual signals or mechanisms. A broader question is how teacher supervision should be allocated when multiple aspects of the interaction jointly determine whether guidance is useful.

We examine this question through three recurring patterns in teacher–student interactions on synthetic WebShop (Figure 1). At the task level, the teacher records no more successes than the student in 62.6% of 2,400 matched task-condition observations with eight rollouts per policy (Figure 1(a)). This Competence Mismatch motivates estimating the teacher's local completion advantage. At the trajectory level, 197 of 301 actions sampled from successful student trajectories receive higher student than teacher conditional scores under identical histories (Figure 1(b)). This Imitation–Outcome Mismatch shows that successful task completion does not imply agreement in action likelihood. At the turn level, conditions with eight teacher successes and seven student successes have mean

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.

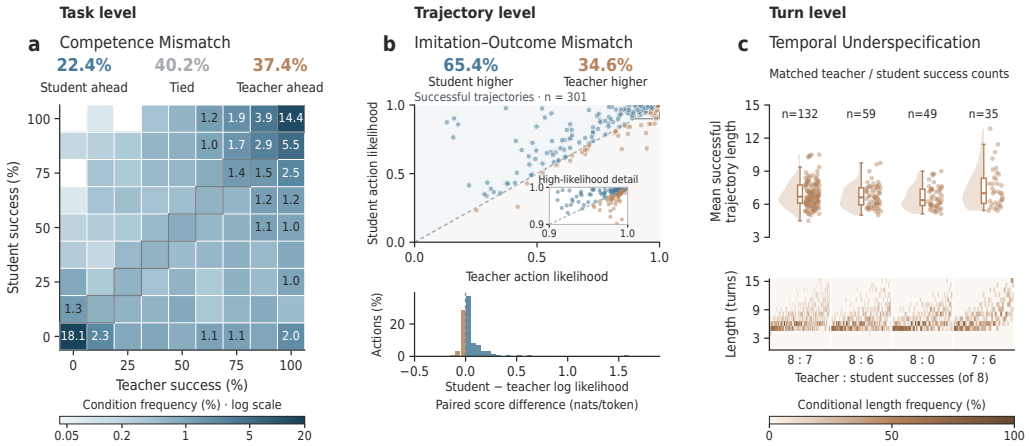

Figure 1: **Empirical motivation for experience-calibrated teacher supervision across task, trajectory, and turn levels in synthetic WebShop (1,000 products).** (a) Task level: matched teacher–student success rates. (b) Trajectory level: conditional action scores on successful student trajectories. (c) Turn level: successful teacher trajectory lengths grouped by matched success counts.

successful teacher lengths ranging from 4.5 to 10.5 turns (Figure 1(c)). This Temporal Underspecification shows that completion counts alone do not capture execution length. These observations motivate supervision weights informed by relative competence, individual outcomes, and temporal completion evidence. Appendix A.1 describes the data sources, sampling, and scoring protocol.

To integrate these three sources of evidence, we propose **Experience-Calibrated On-Policy Distillation (ECOD)**, a unified mechanism for allocating teacher supervision across tasks, trajectories, and turns. Local competence calibration determines task-level supervision strength by comparing teacher and student success-rate posteriors under matched task conditions. Outcome-conditioned selection concentrates direct teacher supervision on failed student trajectories, while all trajectories remain available for reward-based learning. A temporal supervision prior uses successful teacher execution lengths to distribute guidance across turns. These three components form a bounded supervision weight applicable to teacher losses or teacher-derived advantages, allowing the same allocation rule to be integrated into different OPD objectives. Inference uses only the student.

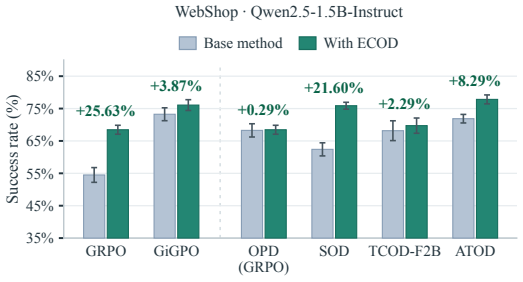

Figure 2: **ECOD on WebShop.** Means ± SD over eight decoding runs; labels show relative gains. GRPO and OPD share the ECOD result.

We evaluate ECOD on ALFWorld and WebShop with Qwen2.5-1.5B-Instruct and Qwen2.5-7B-Instruct, covering GRPO, GiGPO, standard OPD, and three OPD variants (Tables 1 and 2). The reported checkpoint comparisons with the 1.5B student yield relative ALFWorld unseen success gains of 4.06% and 2.56% over GRPO-based and GiGPO-based OPD, respectively, under matched teacher settings. Across both model scales, ECOD-assisted SOD, TCOD-F2B, and ATOD configurations average paired relative gains of 6.20% on ALFWorld unseen and 7.05% on WebShop. Figure 2 summarizes the 1.5B WebShop results. Fixed-endpoint component, teacher-strength, and uniform-allocation controls evaluate supervision allocation separately from the main checkpoint comparisons (Figure 4).

Our contributions are:

- We characterize teacher supervision misalignment in multi-turn OPD at the task, trajectory, and turn levels, and provide empirical evidence motivating unified supervision allocation across these levels.

- We propose ECOD, which jointly calibrates task-level strength, trajectory selection, and turn-level allocation through a bounded supervision weight applicable to teacher-loss and teacher-advantage formulations.
- We evaluate ECOD across RL backbones and OPD variants on ALFWorld and WebShop, and assess the contribution of selective supervision allocation through component ablations and teacher-strength controls.

2 Related Work

2.1 Long-horizon agent reinforcement learning

Research on language agents covers reasoning and feedback during interaction (Yao et al., 2022b; Shinn et al., 2023), instruction tuning, and environment-based policy learning (Zeng et al., 2024; Xi et al., 2025). WebRL (Qi et al., 2025) uses online curricula, AgentGym-RL (Xi et al., 2026) expands interaction horizons, and GiGPO (Feng et al., 2025) improves step-level credit assignment. ECOD adds teacher supervision to reward-based training and computes its weights from completion outcomes and successful teacher execution lengths.

2.2 On-policy distillation

Sequence-level distillation uses teacher-generated outputs (Kim & Rush, 2016), while DAgger (Ross et al., 2011) and OPD (Agarwal et al., 2024) obtain supervision on learner-generated contexts. MiniLLM studies reverse-KL teacher losses (Gu et al., 2024). Agentic methods also incorporate teacher feedback through policy advantages (Tan et al., 2026; Ding et al., 2026). Teacher guidance can be adjusted using divergence, uncertainty, annealing, relative completion rates, and token or trajectory gates (Zhong et al., 2026; Tan et al., 2026; Yu et al., 2026b; Lu et al., 2026; Wu et al., 2026; Li et al., 2026b).

Methods also differ in where they intervene. TCOD and TurnOPD control interaction depth (Wang et al., 2026; Zhou et al., 2026). Guided-OPD mixes teacher and student turns (Li et al., 2026a), while ReOPD and STRIDE select replay or restart positions (Liao et al., 2026; Gui et al., 2026). FutureBridge-OPD and OVCSD assess guidance through future continuations or outcomes (Chen et al., 2026; Xia et al., 2026). OPID, MOPD, and CCOPD construct skill-, peer-, and context-based targets (Yang et al., 2026; Yu et al., 2026a; Lin et al., 2026).

ECOD computes an allocation weight from task-matched completion outcomes and successful teacher lengths. It applies this weight to the teacher contribution while retaining the base method's reward-advantage construction and online student rollout procedure. The allocation rule supports explicit teacher losses and teacher-derived advantages. Appendix A.2 compares the evidence, update terms, and rollout operations of related methods.

3 Preliminaries

3.1 Reinforcement learning for language agents

A language agent π_θ, parameterized by θ, observes history h_t and generates a response x_t containing reasoning and an executable action at turn t. A trajectory is $\tau = (h_t, x_t)_{t=0}^{L-1}$, where L is its length in action turns. Reinforcement learning maximizes expected environment return using reward-derived advantages to update the policy (Schulman et al., 2017). For a task condition u, GRPO samples G trajectories from the rollout policy π_{old} and computes group-relative advantages without a learned critic (Shao et al., 2024). Its clipped policy objective is

$$\mathcal{J}_{\text{GRPO}}^{\text{clip}}(\theta) = \mathbb{E}\Big[\min\Big(r_{k,j}\widehat{A}_k,\ \text{clip}(r_{k,j}, 1-\epsilon, 1+\epsilon)\widehat{A}_k\Big)\Big]. \tag{1}$$

The expectation averages over task conditions and sampled groups, uniformly over trajectories and then response tokens within each trajectory. Here, $r_{k,j} = \pi_\theta(y_{k,j} \mid c_{k,j})/\pi_{\text{old}}(y_{k,j} \mid c_{k,j})$ is the likelihood ratio for response token $y_{k,j}$ given history and response prefix $c_{k,j}$. The advantage is

$\widehat{A}_k = (R_k - \bar{R})/(\sigma_R + \delta)$, where R_k is the trajectory return and $\bar{R}, \sigma_R$ are its group mean and standard deviation. The clipping radius is ϵ, and $\delta > 0$ ensures numerical stability. We denote the negative policy objective together with the base algorithm's regularizers by $\mathcal{L}_{\mathrm{RL}}$.

3.2 On-policy distillation

On-policy distillation (OPD) evaluates a fixed teacher π_T on student-generated contexts, reducing the distribution mismatch associated with fixed demonstrations (Agarwal et al., 2024). Teacher and student predictions condition on the same interaction history and response prefix. With an explicit distillation loss $\mathcal{L}_{\mathrm{distill}}$, the combined training objective is

$$\mathcal{L}(\theta) = \mathcal{L}_{\mathrm{RL}}(\theta) + \lambda\, \mathcal{L}_{\mathrm{distill}}(\theta), \qquad \lambda \geq 0. \tag{2}$$

Here, λ controls teacher supervision strength. Alternatively, teacher feedback can enter the policy update through a teacher-derived advantage. ECOD's weighting of these two forms is defined in Section 4.

4 Experience-Calibrated On-Policy Distillation

ECOD computes teacher-supervision weights from cached teacher rollouts and the current student rollout batch (Figure 3). For each task condition, it compares teacher and student completion outcomes, selects failed student trajectories, and assigns turn weights from successful teacher lengths. These weights are held fixed during optimization and applied to the teacher contribution.

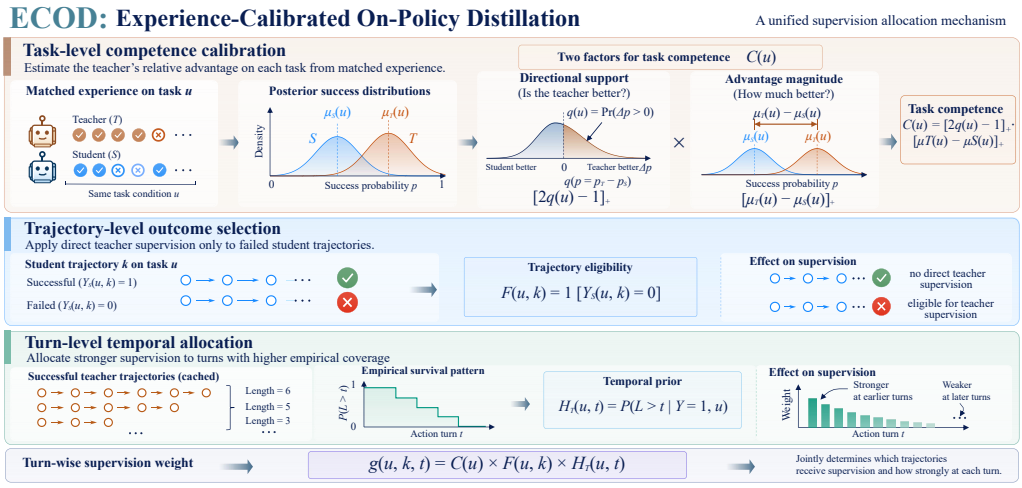

Figure 3: **The framework of ECOD.** Local competence calibration, outcome-conditioned selection, and a temporal supervision prior jointly determine teacher supervision across tasks, trajectories, and turns.

4.1 Experience-calibrated supervision

Let u denote a task and its initial environment configuration, k a student trajectory, and t its zero-based action-turn index. ECOD assigns each turn a supervision weight:

$$g(u,k,t) = \underbrace{C(u)}_{\text{local competence}} \times \underbrace{F(u,k)}_{\text{outcome selection}} \times \underbrace{H_T(u,t)}_{\text{temporal prior}}. \tag{3}$$

Here, $C(u)$ controls task-level strength, $F(u,k)$ selects failed trajectories, and $H_T(u,t)$ allocates supervision across turns. Their joint effect is illustrated at the bottom of Figure 3.

4.1.1 TASK-LEVEL COMPETENCE CALIBRATION

Local competence calibration estimates the teacher's completion advantage from task-matched rollout outcomes (Figure 3, first row).

For task u, records comprise $n_T(u)$ cached teacher trajectories, $\mathcal{D}_T(u) = \{(Y_{T,u,i}, L_{T,u,i})\}_{i=1}^{n_T(u)}$, and $n_S(u)$ current student trajectories, $\mathcal{D}_S(u) = \{Y_{S,u,k}\}_{k=1}^{n_S(u)}$. Here, i indexes teacher trajectories, $Y \in \{0, 1\}$ records failure/success, and L counts action turns. Student statistics use only the current rollout policy π_{old}.

For teacher or student $a \in \{T, S\}$, let p_a be its task-conditioned completion probability under its collection protocol. With conditionally independent Bernoulli outcomes and independent $\text{Beta}(1, 1)$ priors, $s_a = s_a(u)$ successes in $n_a = n_a(u)$ rollouts give the posterior and mean

$$p_a \mid \mathcal{D}_a(u) \sim \text{Beta}\big(s_a + 1, n_a - s_a + 1\big), \qquad \mu_a(u) = \frac{s_a + 1}{n_a + 2}. \tag{4}$$

Let $q(u) = \Pr(p_T > p_S \mid \mathcal{D}_T(u), \mathcal{D}_S(u))$ denote posterior support for a higher teacher completion probability (Figure 3). We combine its directional support with the positive posterior mean gap:

$$C(u) = \underbrace{[2q(u) - 1]_+}_{\text{directional support}} \times \underbrace{[\mu_T(u) - \mu_S(u)]_+}_{\text{advantage magnitude}}, \qquad [z]_+ = \max(z, 0). \tag{5}$$

The first factor measures directional support, and the mean gap measures advantage magnitude. Their product reduces supervision when either is small.

For TCOD-F2B, completion is measured at the teacher's full horizon and the student's current curriculum horizon (Appendix A.3).

4.1.2 TRAJECTORY-LEVEL OUTCOME SELECTION

Outcome-conditioned selection uses task completion to determine eligibility for direct teacher supervision (Figure 3, second row):

$$F(u, k) = \mathbf{1}[Y_{S,u,k} = 0] = 1 - Y_{S,u,k}. \tag{6}$$

Here, $\mathbf{1}[\cdot]$ is the indicator function. Successful trajectories receive zero direct teacher weight and remain in the reward-based update. Failed trajectories are eligible for supervision, with weights further determined by competence and temporal coverage. All trajectories contribute to estimating $C(u)$.

4.1.3 TURN-LEVEL TEMPORAL ALLOCATION

The temporal supervision prior uses successful teacher lengths to allocate guidance across action turns:

$$H_T(u, t) = \begin{cases} \dfrac{1}{s_T(u)} \displaystyle\sum_{i=1}^{n_T(u)} Y_{T,u,i}\, \mathbf{1}[L_{T,u,i} > t], & s_T(u) > 0, \\ 0, & s_T(u) = 0. \end{cases} \tag{7}$$

For tasks with successful teacher rollouts, the temporal prior is the empirical fraction of successful teacher trajectories that contain action turn t. It measures task-conditioned coverage over execution length and is applied within the base student rollout. Each factor lies in $[0, 1]$, so their product also lies in $[0, 1]$. ECOD composes this weight with the base method's configured teacher weighting and schedule.

4.2 POLICY OPTIMIZATION WITH ECOD

ECOD uses the same allocation weight in teacher-loss and teacher-advantage formulations. Each formulation retains its policy-update computation and reward-advantage construction. We denote stop-gradient by $\mathrm{sg}(\cdot)$.

Teacher-loss weighting. For OPD and TCOD-F2B, ECOD weights the teacher loss $\ell^T_{u,k,t,j}(\theta)$ for response token j before aggregation:

$$\mathcal{L}_{\mathrm{ECOD}}(\theta) = \mathcal{L}_{\mathrm{RL}}(\theta) + \frac{\lambda}{N} \sum_{(u,k,t,j)\in\mathcal{I}} \mathrm{sg}[g(u,k,t)]\ \ell^T_{u,k,t,j}(\theta), \tag{8}$$

Here, $\mathcal{L}_{\mathrm{RL}}$ includes the reward objective and existing regularizers, and $\lambda \geq 0$ is the base teacher coefficient. The set $\mathcal{I}$ contains all valid response-token positions in a microbatch, with $N = |\mathcal{I}|$, including zero-weight positions. Averaging over these tokens lets ECOD control both the distribution and the overall strength of supervision.

Teacher-advantage weighting. In our shared GRPO implementation, SOD and ATOD incorporate teacher feedback through a teacher-derived advantage. ECOD weights this contribution before combining it with the reward-derived advantage:

$$\widehat{A}^{\mathrm{ECOD}}_{u,k,t,j} = \widehat{A}^{\mathrm{RL}}_{u,k,t,j} + \mathrm{sg}[g(u,k,t)]\ \widehat{A}^{T}_{u,k,t,j}, \tag{9}$$

The reward and teacher terms use the coefficients configured for each experimental setting. The base policy objective is evaluated with the combined advantage. Appendix A.3 gives the implementation details, including the SOD adaptation; Algorithm 1 in Appendix A.4 gives the training procedure.

5 EXPERIMENTS

We evaluate ECOD through comparisons across RL and OPD methods, controlled allocation ablations, and training dynamics. These analyses assess integration across teacher objectives, selective allocation under matched supervision budgets, and teacher weighting during learning.

5.1 EXPERIMENTAL SETUP

Benchmarks. We use ALFWorld for household task completion (Shridhar et al., 2020) and WebShop for instruction-guided shopping (Yao et al., 2022a). WebShop uses the 1,000-product configuration with synthetic goals. Appendix A.5 specifies the task splits, reset seeds, and interaction limits.

Baselines. We compare GRPO and GiGPO with their ECOD-assisted counterparts (Shao et al., 2024; Feng et al., 2025). We evaluate standard OPD on both backbones and use shared-framework implementations of SOD, TCOD-F2B, and ATOD in the multi-turn GRPO framework (Zhong et al., 2026; Wang et al., 2026; Tan et al., 2026). Each teacher-assisted method is evaluated with and without ECOD. Appendix A.6 specifies the baseline implementations.

Students and teachers. Students are initialized from Qwen2.5-1.5B-Instruct or Qwen2.5-7B-Instruct and train for 150 updates (Qwen et al., 2024). Paired teacher-assisted configurations share a frozen teacher of the same model size, trained for 150 updates with the corresponding benchmark and RL backbone. On 7B WebShop, ECOD-assisted runs use the teacher strengths specified in Appendix A.7; paired baselines retain their original strengths.

Evaluation. Validation is performed every five updates on 128 tasks. Main results use the designated checkpoint for each configuration, and auxiliary metrics use the same checkpoint. Ablations use update 150. Each configuration is trained once. Except for the seen entries detailed in Appendix A.8, reported means and sample standard deviations use eight decoding runs at fixed model weights. All evaluations use the student alone.

Table 1: Main results with Qwen2.5-1.5B-Instruct.

		ALFWorld		WebShop	
Baseline	Configuration	Seen SR (%) ↑	Unseen SR (%) ↑	SR (%) ↑	Task score (%) ↑
GRPO	Base	61.43	57.56±1.23	54.49±2.28	64.46±1.87
	+ ECOD	75.00 ↑ **22.09**	74.16±0.97 ↑ **28.85**	68.46±1.38 ↑ **25.63**	80.75±1.71 ↑ **25.27**
GiGPO	Base	82.86	83.58±0.00	73.24±1.99	86.74±1.57
	+ ECOD	**92.86** ↑ **12.07**	**89.55**±0.00 ↑ **7.14**	76.07±1.67 ↑ **3.87**	**89.42**±0.85 ↑ **3.09**
OPD	GRPO	75.00	71.27±2.22	68.26±2.04	77.39±1.95
	GRPO + ECOD	75.00 ↔ 0.00	74.16±0.97 ↑ **4.06**	68.46±1.38 ↑ **0.29**	80.75±1.71 ↑ **4.35**
	GiGPO	**92.86**	87.31±0.00	76.07±1.32	89.31±1.12
	GiGPO + ECOD	**92.86** ↔ 0.00	**89.55**±0.00 ↑ **2.56**	76.07±1.67 ↔ 0.00	**89.42**±0.85 ↑ **0.12**
SOD	Base	65.00	66.98±1.25	62.40±2.02	74.27±2.81
	+ ECOD	75.71 ↑ **16.48**	81.34±0.00 ↑ **21.45**	75.88±1.06 ↑ **21.60**	88.23±0.87 ↑ **18.80**
TCOD	F2B	70.00	70.34±1.25	68.16±3.06	79.06±2.63
	F2B + ECOD	80.00 ↑ **14.29**	79.10±0.98 ↑ **12.47**	69.73±2.35 ↑ **2.29**	82.03±1.86 ↑ **3.76**
ATOD	Base	74.29	78.64±1.05	71.88±1.32	84.59±1.06
	+ ECOD	80.71 ↑ **8.65**	73.51±1.06 ↓ **6.52**	**77.83**±1.38 ↑ **8.29**	87.89±0.88 ↑ **3.89**

Table 2: Main results with Qwen2.5-7B-Instruct.

		ALFWorld		WebShop	
Baseline	Configuration	Seen SR (%) ↑	Unseen SR (%) ↑	SR (%) ↑	Task score (%) ↑
GRPO	Base	77.14	72.39±0.00	73.34±3.67	83.10±2.64
	+ ECOD	77.14 ↔ 0.00	72.39±0.00 ↔ 0.00	75.39±1.32 ↑ **2.80**	86.05±1.04 ↑ **3.55**
GiGPO	Base	89.29	87.50±0.66	74.51±1.38	84.56±1.96
	+ ECOD	**95.71** ↑ **7.20**	**91.32**±0.79 ↑ **4.37**	**82.71**±0.88 ↑ **11.01**	**92.97**±0.72 ↑ **9.94**
OPD	GRPO	75.71	72.39±2.15	77.05±1.32	85.91±1.11
	GRPO + ECOD	77.14 ↑ **1.89**	72.39±0.00 ↔ 0.00	75.39±1.32 ↓ **2.15**	86.05±1.04 ↑ **0.17**
	GiGPO	93.57	90.11±1.31	80.47±1.02	88.99±0.90
	GiGPO + ECOD	**95.71** ↑ **2.29**	**91.32**±0.79 ↑ **1.35**	**82.71**±0.88 ↑ **2.79**	**92.97**±0.72 ↑ **4.46**
SOD	Base	79.29	75.56±1.04	79.88±2.04	87.50±1.14
	+ ECOD	82.86 ↑ **4.50**	75.37±0.00 ↓ **0.25**	82.03±1.82 ↑ **2.69**	90.28±1.26 ↑ **3.18**
TCOD	F2B	81.43	74.63±0.00	77.54±1.30	86.76±1.18
	F2B + ECOD	84.29 ↑ **3.51**	81.34±1.60 ↑ **9.00**	79.88±0.91 ↑ **3.02**	88.47±1.01 ↑ **1.98**
ATOD	Base	83.57	79.10±0.00	79.20±1.72	87.15±1.40
	+ ECOD	85.71 ↑ **2.56**	79.94±1.93 ↑ **1.06**	**82.71**±0.77 ↑ **4.44**	89.84±0.87 ↑ **3.08**

SR: success rate; task score: normalized reward, both expressed as percentages. Each row reports the designated checkpoint for that configuration. Unseen and WebShop metrics report means and sample standard deviations over eight decoding runs at fixed weights. Seen entries omit standard deviations; their repeat counts are specified in Appendix A.8. Arrows give relative changes from paired baselines using unrounded means. Bold marks the best entry in each column. The 7B WebShop ECOD configurations use the adjusted teacher strengths in Section 5.1. SOD denotes the shared-framework implementation in Appendix A.6.

5.2 Main results

Tables 1 and 2 compare the reported test results of each method and its ECOD-assisted counterpart.

Allocation beyond standard distillation. With the 1.5B student on ALFWorld unseen, GRPO-based OPD records 71.27% success and its ECOD counterpart records 74.16%. GiGPO-based OPD records 87.31% and its ECOD counterpart records 89.55%. These correspond to relative gains of 4.06% and 2.56%. Each pair shares the teacher and base teacher coefficient. Section 5.3 separately evaluates allocation at a fixed training endpoint.

Integration across distillation formulations. The ECOD-assisted SOD, TCOD-F2B, and ATOD configurations have higher reported success in all six WebShop comparisons, with relative gains ranging from 2.29% to 21.60%. Across the two model scales, the mean paired relative changes are

6.20% on ALFWorld unseen and 7.05% on WebShop. These comparisons cover both teacher-loss and teacher-advantage implementations. Tables 1 and 2 report every paired result.

Comparison with reward-based training. With the 1.5B student, ECOD-assisted GRPO and GiGPO have higher reported success than their RL-only counterparts on both benchmarks. With the 7B student, GiGPO with ECOD records 91.32% on ALFWorld unseen and 82.71% on WebShop. These comparisons evaluate teacher-assisted training alongside GiGPO's step-level credit assignment.

5.3 Ablation studies

We evaluate allocation factors and teacher-strength controls with Qwen2.5-1.5B-Instruct and GRPO at update 150 (Figure 4). Each WebShop control matches the weighted-token total that full ECOD would assign to the control's current rollout batch. It redistributes this total using an allocation rule with one factor removed, or with a pooled temporal prior.

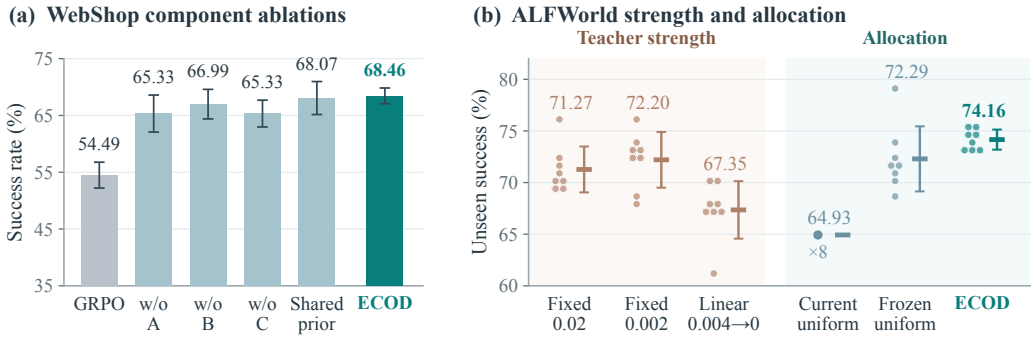

Figure 4: **Allocation ablations.** (a) w/o A/B/C removes competence weighting/outcome selection/temporal weighting from the allocation pattern, retaining the ECOD-derived weighted-token total on each control's batch. (b) ALFWorld strength and uniform-allocation controls (Current/Frozen: current-batch/frozen budgets). Mean ± sample s.d. over eight decoding runs at step 150; dots show individual runs.

Allocation factors. Full ECOD reaches 68.46% mean WebShop success. Removing competence weighting, outcome selection, or temporal weighting from the allocation pattern gives 65.33%, 66.99%, and 65.33%, respectively. A pooled temporal prior gives 68.07%. Appendix A.9 defines the matching rule.

Teacher strength and uniform allocation. On ALFWorld unseen, full ECOD reaches 74.16% mean success. Uniform OPD coefficients of 0.02 and 0.002 give 71.27% and 72.20%, while linear annealing gives 67.35%. Current-batch and frozen-budget uniform allocation give 64.93% and 72.29%, respectively. Full ECOD has the highest mean success among these fixed-endpoint controls. Appendix A.10 specifies the coefficient schedules and budget-matching procedures.

5.4 Teacher weighting during training

To examine learning dynamics, we compare teacher coefficients and validation success for SOD, TCOD-F2B, and ATOD on WebShop (Figure 5).

Over updates 126–150, mean teacher coefficients under ECOD are 1.09%, 2.08%, and 1.38% of the paired SOD, TCOD-F2B, and ATOD baselines. Mean success over the final five validation events is higher by 15.01%, 2.79%, and 6.18%, respectively. The curves show higher late-stage success alongside lower teacher weighting. Appendix A.11 defines the averaging protocol. Appendix A.12 reports active-turn coverage and nonzero-weight distributions for the GRPO-based runs.

6 Conclusion

ECOD allocates teacher supervision from task-matched completion outcomes and successful teacher execution lengths. It combines task-level competence calibration, trajectory-level outcome selection, and turn-level temporal allocation in a weight applied to teacher losses or teacher-derived

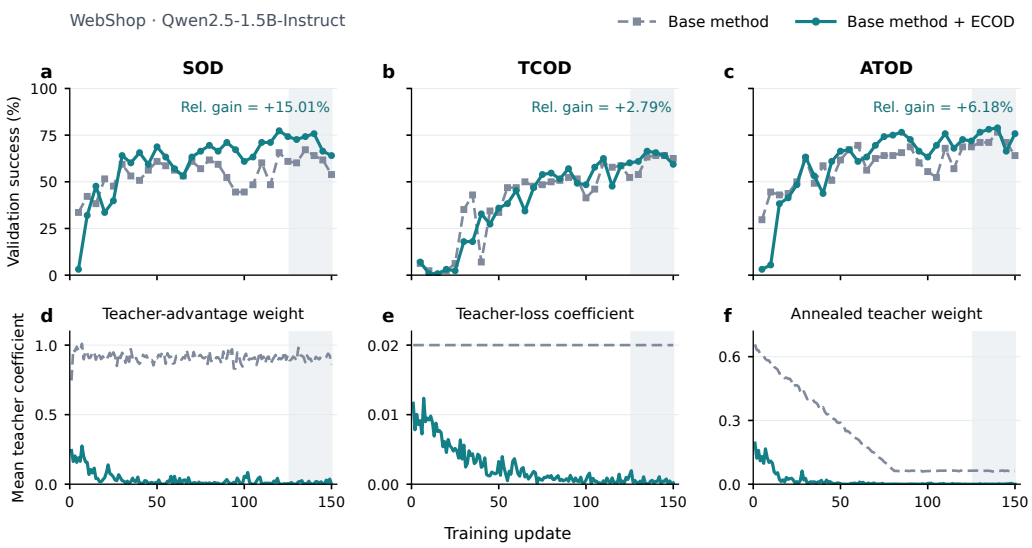

Figure 5: **Learning dynamics on WebShop (1.5B, GRPO).** Dashed/solid curves denote base/ECOD methods: (a–c) validation success; (d–f) mean teacher coefficients on method-specific scales. Shading marks updates 126–150; annotations report relative gains in mean success. Evaluation follows Section 5.1; curve statistics are detailed in Appendix A.11.

advantages. Experiments on ALFWorld and WebShop cover two model scales and multiple RL and OPD implementations. Fixed-endpoint controls show higher mean completion success for full ECOD than for the tested coefficient schedules and uniform-allocation rules. The trained student acts without teacher intervention at inference.

AI Use Statement

In this work, we used generative AI tools to assist with literature search and background research, as well as language editing and stylistic refinement of author-written manuscript drafts. We also used generative AI tools to assist with portions of coding and the implementation and refinement of the experimental pipeline. The research ideas, methodological design, experimental procedures, selection of benchmarks, baselines, datasets, and models, as well as the initial experimental framework, were developed by the authors. AI tools were used primarily to help translate these designs into code and improve their implementation. All AI-assisted text, code, and experimental components were reviewed, revised, and verified by the authors. We take full responsibility for the final content and results of this work.

REFERENCES

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

Chishui Chen, Yaoyou Fan, Te Sun, Yi Yang, Chenghao Sun, Delin Mao, Hongbo Qiao, Zuowei Zhang, Junxi Wang, Chenxing Sun, Yangen Hu, Lu Pan, Xuyang Liu, and Linfeng Zhang. Look Ahead Before You Distill: Future Trajectory Validation of Teacher Guidance for Agentic On-Policy Distillation. *arXiv preprint arXiv:2608.01953*, 2026. URL https://arxiv.org/abs/2608.01953.

Ken Ding. HDPO: Hybrid Distillation Policy Optimization via Privileged Self-Distillation. *arXiv preprint arXiv:2603.23871*, 2026. URL https://arxiv.org/abs/2603.23871.

Yifan Ding, Xincheng Wei, Yoshua Y. Li, Ziheng Li, Yuquan Lu, Siyu Zhang, Dongsheng Ma, Rongxiang Weng, Xunliang Cai, and Yun Chen. SAF-OPD: Stable Advantage Fusion for On-Policy Distillation. *arXiv preprint arXiv:2607.29209*, 2026. URL https://arxiv.org/abs/2607.29209.

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, pp. 46375–46408, 2025. doi: 10.52202/085713-1544. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/420c9f777c0b4f78d515e53cf74d58b2-Abstract-Conference.html.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiacai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting On-Policy Distillation: Empirical Failure Modes and Simple Fixes. *arXiv preprint arXiv:2603.25562*, 2026. URL https://arxiv.org/abs/2603.25562.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024.

Zhiyu Gui, Kexin Huang, Jia Guo, Junkang Wu, Zihao Wang, Zhiqiang Zhang, Jun Zhou, Jiancan Wu, and Xiang Wang. Know When to Stop, Where to Restart: Accelerating Multi-Turn Agentic On-Policy Distillation. *arXiv preprint arXiv:2609.14636*, 2026. URL https://arxiv.org/abs/2609.14636.

Jeonghye Kim, Jiwon Jeon, Dongsheng Li, and Yuqing Yang. Rebellious Student: Reversing Teacher Signals for Reasoning Exploration with Self-Distilled RLVR. *arXiv preprint arXiv:2605.10781*, 2026. URL https://arxiv.org/abs/2605.10781.

Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 1317–1327, 2016.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

Gengsheng Li, Mao Zheng, Mingyang Song, Ruiqi Liu, Tianyu Yang, Jie Sun, Qiyong Zhong, Haiyun Guo, Junfeng Fang, Dan Zhang, and Jinqiao Wang. On-Policy Distillation with Curriculum Turn-level Guidance for Multi-turn Agents. *arXiv preprint arXiv:2606.15912*, 2026a. URL https://arxiv.org/abs/2606.15912.

Xiaozhe Li, Tianyi Lyu, Yang Li, Yichuan Ma, Peiji Li, Linyang Li, Qipeng Guo, Dahua Lin, and Kai Chen. What and When to Distill: Selective Hindsight Distillation for Multi-Turn Agents. *arXiv preprint arXiv:2605.19447*, 2026b. URL https://arxiv.org/abs/2605.19447.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking On-Policy Distillation of Large Language Models: Phenomenology, Mechanism, and Recipe. *arXiv preprint arXiv:2604.13016*, 2026c. URL https://arxiv.org/abs/2604.13016.

Baohao Liao, Hanze Dong, Christof Monz, Xinxing Xu, Li Dong, and Furu Wei. Multi-Turn On-Policy Distillation with Prefix Replay. *arXiv preprint arXiv:2607.04763*, 2026. URL https://arxiv.org/abs/2607.04763.

Zizhuo Lin, Quanling Liu, Jinsheng Quan, Chao Zhang, Yifan Zhu, Xing Shi, Jingtao Xu, Zhihui Li, and Yawei Luo. Same Evidence, Different Answers: Canonical-Context On-Policy Distillation for Multi-Turn Language Models. *arXiv preprint arXiv:2605.30251*, 2026. URL https://arxiv.org/abs/2605.30251.

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

Zehan Qi, Xiao Liu, Iat Long Iong, Hanyu Lai, Xueqiao Sun, Jiadai Sun, Xinyue Yang, Yu Yang, Shuntian Yao, Wei Xu, Jie Tang, and Yuxiao Dong. WebRL: Training LLM Web Agents via Self-Evolving Online Curriculum Reinforcement Learning. In *International Conference on Learning Representations*, pp. 79791–79821, 2025.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635. JMLR Workshop and Conference Proceedings, 2011.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

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

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

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*, 2020.

Mingyang Song and Mao Zheng. A Survey of On-Policy Distillation for Large Language Models. *arXiv preprint arXiv:2604.00626*, 2026. URL https://arxiv.org/abs/2604.00626.

Qitai Tan, Zefang Zong, Mo Li, Yipeng Shi, Yang Li, and Peng Chen. ATOD: Annealed Turn-Aware On-Policy Distillation for Multi-Turn Agentic Tasks. *arXiv preprint arXiv:2606.27814*, 2026. URL https://arxiv.org/abs/2606.27814.

Jiaqi Wang, Wenhao Zhang, Weijie Shi, Yaliang Li, and James Cheng. TCOD: Exploring Temporal Curriculum in On-Policy Distillation for Multi-turn Autonomous Agents. *arXiv preprint arXiv:2604.24005*, 2026. URL https://arxiv.org/abs/2604.24005.

Jinyang Wu, Shuo Yang, Zhengxi Lu, Fan Zhang, Yuhao Shen, Lang Feng, Haoran Luo, Zheng Lian, Shuai Zhang, Zhengqi Wen, and Jianhua Tao. SEED: Self-Evolving On-Policy Distillation for Agentic Reinforcement Learning. *arXiv preprint arXiv:2607.14777*, 2026. URL https://arxiv.org/abs/2607.14777.

Zhiheng Xi, Yiwen Ding, Wenxiang Chen, Boyang Hong, Honglin Guo, Junzhe Wang, Xin Guo, Dingwen Yang, Chenyang Liao, Wei He, Songyang Gao, Lu Chen, Rui Zheng, Yicheng Zou, Tao Gui, Qi Zhang, Xipeng Qiu, Xuanjing Huang, Zuxuan Wu, and Yu-Gang Jiang. AgentGym: Evaluating and Training Large Language Model-based Agents across Diverse Environments. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27914–27961. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1355.

Zhiheng Xi, Jixuan Huang, Chenyang Liao, Baodai Huang, Jiaqi Liu, Honglin Guo, Yajie Yang, 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: An open-source framework to train llm agents for long-horizon decision making via multi-turn rl. In *International Conference on Learning Representations*, pp. 12685–12727, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/1571ce1ff735be4551db50887043726e-Abstract-Conference.html.

Xu Xia, Jinghua Piao, Min Yang, Xiaochong Lan, Jiaju Chen, and Yong Li. From Scoring to Acting: Outcome-Verified Comparative Self-Distillation for LLM Agents. *arXiv preprint arXiv:2607.27937*, 2026. URL https://arxiv.org/abs/2607.27937.

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. *arXiv preprint arXiv:2606.26790*, 2026. URL https://arxiv.org/abs/2606.26790.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35:20744–20757, 2022a.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022b.

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. *arXiv preprint arXiv:2605.12652*, 2026a. URL https://arxiv.org/abs/2605.12652.

Yan Yu, Zhengxi Lu, Yizhou Liu, Yichen Pan, Aozhe Wang, Qipeng Chen, Hua Yang, Wenqi Zhang, Weiming Lu, Qianglong Chen, and Yongliang Shen. RetireOPD: Self-Retiring On-Policy Distillation for Agentic Reinforcement Learning. *arXiv preprint arXiv:2609.20784*, 2026b. URL https://arxiv.org/abs/2609.20784.

Aohan Zeng, Mingdao Liu, Rui Lu, Bowen Wang, Xiao Liu, Yuxiao Dong, and Jie Tang. AgentTuning: Enabling Generalized Agent Abilities for LLMs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3053–3077, 2024.

Qiyong Zhong, Mao Zheng, Mingyang Song, Xin Lin, Jie Sun, Houcheng Jiang, Xiang Wang, and Junfeng Fang. SOD: Step-wise On-policy Distillation for Small Language Model Agents. *arXiv preprint arXiv:2605.07725*, 2026. URL https://arxiv.org/abs/2605.07725.

Yuhang Zhou, Kai Zheng, Haoling Li, Dengyun Peng, Can Xu, and Jingjing Chen. TurnOPD: Making On-Policy Distillation Turn-Aware for Efficient Long-Horizon Agent Training. *arXiv preprint arXiv:2607.05804*, 2026. URL https://arxiv.org/abs/2607.05804.

A SUPPLEMENTARY MATERIAL

Appendix contents

A.1 Motivation analysis protocol . . . 13
A.2 Comparison with recent OPD methods . . . 13
A.3 ECOD implementation . . . 14
A.4 Training algorithm . . . 16
A.5 Benchmark configurations . . . 16
A.6 Baseline implementations . . . 16
A.7 Training details . . . 17
A.8 Evaluation details . . . 18
A.9 Matched-allocation ablations . . . 19
A.10 Teacher-strength and allocation controls . . . 19
A.11 Learning-dynamics protocol . . . 20
A.12 Supervision coverage and component distributions . . . 20

A.1 MOTIVATION ANALYSIS PROTOCOL

Figure 1 uses Qwen2.5-1.5B-Instruct students in synthetic WebShop with 1,000 products and a 15-turn limit. The fixed teacher is the corresponding GRPO policy after 150 updates.

Task and turn panels. Panels (a) and (c) use one GRPO-based OPD run over updates 1–150. Each update contains 16 task conditions, with eight student rollouts and eight cached teacher rollouts per condition. The 2,400 entries are task-condition observations across training and may include repeated tasks. Panel (c) groups entries by teacher and student success counts. Among groups with at least four teacher successes and more teacher than student successes, it shows the four largest groups, containing 132, 59, 49, and 35 conditions. In the 8:7 group, each point is the mean of eight successful teacher lengths for one condition; these means range from 4.5 to 10.5 turns. Heatmap columns show condition-specific length distributions normalized by the number of successful teacher trajectories.

Action-score panel. Panel (b) pools successful trajectories from GRPO, OPD, and SOD at checkpoint 125, using rollout update 126, and TCOD-F2B at checkpoint 145, using rollout update 146. The four 128-trajectory batches contain 303 successful trajectories. A fixed SHA-256 rule selects one action turn per trajectory. Two OPD responses without a unique nonempty action span are excluded without replacement. The remaining 301 actions comprise 73 GRPO, 69 OPD, 71 SOD, and 88 TCOD-F2B samples.

The generating student checkpoint and the fixed teacher score identical action tokens under the recorded history and response prefix, including preceding reasoning. Scores are mean log probabilities over action-content tokens. The scatterplot shows exponentiated scores, and the histogram shows student-minus-teacher differences in nats per token. Each action has equal weight.

A.2 COMPARISON WITH RECENT OPD METHODS

Table A1 distinguishes the evidence used for supervision, the affected update term, and changes to online student rollouts.

Table A1: Comparison of supervision evidence and intervention mechanisms.

Method	Supervision evidence: Completion evidence and scope	Supervision evidence: Outcome	Supervision evidence: Turn	Optimization: Teacher term	Sampling: Online rollout retained
Direct experimental comparators					
SOD (Zhong et al., 2026)	–	–	✓	Loss	✓
TCOD-F2B (Wang et al., 2026)	–	–	△	Loss	–
ATOD (Tan et al., 2026)	–	–	✓	Teacher advantage	✓
Adaptive strength and gating					
RetireOPD (Yu et al., 2026b)	Global T–S rates	△	–	Loss	✓
SDAR (Lu et al., 2026)	–	△	–	Loss	✓
SEED (Wu et al., 2026)	Completed trajectories	△	–	Loss	✓
SAF-OPD (Ding et al., 2026)	–	–	–	Teacher advantage	✓
SERL (Li et al., 2026b)	Reward-advantage sign	△	✓	Loss + RL advantage	✓
Turn and temporal allocation					
Guided-OPD (Li et al., 2026a)	–	–	△	Loss	–
TurnOPD (Zhou et al., 2026)	Student probe successes	△	✓	Loss	–
ReOPD (Liao et al., 2026)	–	–	✓	Loss	–
STRIDE (Gui et al., 2026)	–	–	✓	Loss	–
Outcome selection and future validation					
FutureBridge-OPD (Chen et al., 2026)	–	△	△	Teacher advantage	–
OVCSD (Xia et al., 2026)	Group + repair outcomes	✓	△	Loss + comparative RL	–
HDPO (Ding, 2026)	Group + solution outcomes	✓	–	Loss	–
RLRT (Kim et al., 2026)	Student correctness	✓	–	Reward advantage	✓
Peer, skill, and context construction					
MOPD (Yu et al., 2026a)	Peer rollout outcomes	✓	–	Loss	✓
OPID (Yang et al., 2026)	Completed trajectories	△	△	Teacher advantage	✓
CCOPD (Lin et al., 2026)	–	–	△	Loss	✓
ECOD (ours)	Task-conditioned T–S counts	✓	✓	Loss or teacher advantage	✓

✓ explicit mechanism; △ related mechanism with different granularity; – absent from the allocation or target-construction rule. Completion evidence excludes ordinary RL rewards and specifies how outcomes enter the teacher-related rule (T–S: teacher–student). Outcome denotes trajectory selection; turn denotes within-trajectory allocation. The teacher-term column follows original formulations. SOD specifies a weighted OPD loss; our shared-framework implementation applies its step-weighting mechanism through a teacher-derived advantage. Online rollout retained refers to the base online student rollout procedure. ECOD separately precollects teacher trajectories for allocation statistics.

RetireOPD uses global teacher-student completion rates to determine teacher retirement. ECOD recomputes task-conditioned completion statistics and weights individual failed trajectories. TurnOPD uses successful student lengths and KL statistics to control rollout depth; ECOD uses successful teacher lengths to weight turns within retained student rollouts. OVCSD and HDPO select all-failure groups and verify teacher continuations or solutions. ECOD's trajectory gate also selects failures within mixed-outcome groups. Table A1 records the evidence and intervention locations of the other methods.

Related analyses examine token-support matching (Fu et al., 2026), teacher-student compatibility and prompt selection (Li et al., 2026c), and the broader design space of OPD objectives and training dynamics (Song & Zheng, 2026).

A.3 ECOD IMPLEMENTATION

Each turn weight $g(u, k, t)$ applies to all valid reasoning and action tokens in that response, excluding environment observations and padding. Teacher and student predictions share the same interaction history and response prefix. Allocation weights are detached during policy optimization.

A.3.1 ROLLOUT STATISTICS AND POSTERIOR COMPARISON

Teacher statistics are collected once from the frozen teacher. Student statistics are recomputed from each new rollout batch under the rollout policy π_{old}. Each trajectory contributes one binary completion outcome. Completion is assessed at each policy's collection horizon; for TCOD-F2B, the teacher uses the full horizon and the student uses its current curriculum horizon (Appendix A.7).

Let $a_T = s_T + 1$, $b_T = n_T - s_T + 1$, $a_S = s_S + 1$, and $b_S = n_S - s_S + 1$. The integer Beta shapes permit deterministic evaluation of the posterior comparison:

$$q(u) = \sum_{i=0}^{a_T - 1} \frac{B(a_S + i, b_T + b_S)}{(b_T + i) B(1 + i, b_T) B(a_S, b_S)}. \tag{10}$$

Here, B is the beta function. The implementation evaluates each summand through log-beta values computed with log-gamma functions, caches repeated count combinations, and clips numerical results to $[0, 1]$. The temporal table stores the empirical survival fraction for every required task–turn pair, including genuine zeros. Missing entries are treated as configuration errors rather than zero temporal support.

A.3.2 Explicit teacher loss

The explicit teacher loss uses the importance-corrected sampled surrogate. For each sampled response token, let z_θ, z_{old}, and z_T denote the current-student, rollout-student, and teacher log probabilities on the same context. Teacher and rollout-student scores are detached. The implemented surrogate is

$$\begin{aligned} d &= \text{clip}(z_\theta - z_T, -20, 20), & r &= \exp(\text{clip}(z_\theta - z_{\text{old}}, -20, 20)), \\ v &= \text{clip}\big(r(e^{-d} - 1 + d), 0, 10\big), & \ell^T &= \text{sg}(v) + \text{sg}(rd)\,(z_\theta - \text{sg}(z_\theta)). \end{aligned} \tag{11}$$

The second term has zero forward value and supplies the gradient $\text{sg}(rd)$ times the current-student log-probability gradient. Without clipping, this equals the gradient of the importance-weighted sampled expression with the sampled context held fixed. The forward cap applies to v; the detached gradient coefficient uses the clipped log ratios. ECOD multiplies the surrogate by its detached turn weight before reduction.

A.3.3 Reduction and parameter updates

The explicit teacher term uses `all_response_token_mean` within each device-local microbatch. Its denominator counts all valid response tokens, including zero-weight positions, and excludes environment observations and padding. A microbatch with zero total supervision contributes zero teacher loss. Training averages these microbatch objectives across gradient-accumulation steps and retains the base FSDP gradient reduction. Microbatch means receive equal accumulation weights, independently of their token counts. The same reduction is used by each paired explicit-loss baseline and its ECOD extension.

A.3.4 Teacher-derived advantages

For SOD and ATOD, the teacher token signal is $\Delta_j = z_{T,j} - z_{\text{old},j}$, computed once per rollout batch with detached scores. Let w_t denote the base method's turn weight. SOD obtains w_t from its divergence-ratio rule; ATOD computes it from within-trajectory normalized disagreement and uncertainty. ECOD acts after these native weights have been constructed:

$$\begin{aligned} \widehat{A}^{\text{SOD+ECOD}}_{t,j} &= \widehat{A}^{\text{RL}}_{t,j} + g_t\, \eta w_t \Delta_j, \\ \widehat{A}^{\text{ATOD+ECOD}}_{t,j} &= \rho_m \widehat{A}^{\text{RL}}_{t,j} + g_t\, \kappa_m w_t \Delta_j. \end{aligned} \tag{12}$$

Task and trajectory indices are suppressed. Here, η is the configured SOD teacher strength, and ρ_m, κ_m follow the ATOD schedule at update m. The 1.5B SOD implementation uses $\eta = 1$; the 7B WebShop adjustment is specified in Appendix A.7. Reward-derived advantages are computed and normalized by the base method before this combination. There is no additional normalization of the combined advantage before the actor's PPO objective. ECOD preserves the construction and normalization of reward-derived advantages and retains native schedules; PPO clipping is subsequently applied to the combined advantage. All advantages and ECOD weights remain fixed during optimization of the rollout batch.

A.4 TRAINING ALGORITHM

Algorithm 1 Training with ECOD

Inputs: student π_θ, fixed teacher π_T, training task conditions, base policy optimizer with a separable teacher contribution
Prepare: collect teacher rollouts; cache success counts and successful lengths
For each training update:
1: Set $\pi_{\text{old}} \leftarrow \pi_\theta$; collect student rollouts for the sampled task conditions.
2: Record one completion outcome per rollout; compute student counts.
3: Compare teacher/student posteriors to obtain $C(u)$.
4: Construct $F(u, k)$ from student outcomes and query $H_T(u, t)$.
5: Form $g(u, k, t) = C(u) \times F(u, k) \times H_T(u, t)$; detach the coefficients.
6: Evaluate teacher feedback on the student-generated token contexts.
7: Weight the teacher loss [Eq. (8)] or teacher advantage [Eq. (9)] before aggregation or combination with the reward-derived term.
8: Update θ with the base optimizer, keeping g fixed for this batch.
Output: student policy π_θ; no teacher is required at inference.

Cached teacher rollouts provide allocation statistics; online teacher scoring supplies the token-level training signal.

A.5 BENCHMARK CONFIGURATIONS

ALFWorld. We used ALFWorld's text-based TextWorld environment (Shridhar et al., 2020). Following the environment filtering in our verl-agent implementation, the task pools contained 3,553 training games, 140 seen evaluation games, and 134 unseen evaluation games. We included all six task types: Pick, Look, Clean, Heat, Cool, and Pick2. The interaction horizon was capped at 50 action turns, with the two most recent observation–action turns retained in the history.

WebShop. We used the 1,000-product, synthetic-goal WebShop configuration in verl-agent (Yao et al., 2022a). Training and evaluation reset seeds were sampled from the disjoint ranges $[0, 2^{16})$ and $[2^{16}, 2^{32})$, respectively. Final evaluation used a fixed schedule of 128 task resets. We allowed at most 15 action turns and retained the two most recent observation–action turns.

A.6 BASELINE IMPLEMENTATIONS

We organized the baselines into two groups. GRPO and GiGPO provide RL-only reference points, whereas OPD, SOD, TCOD, and ATOD examine different uses of teacher supervision.

- **GRPO:** Group Relative Policy Optimization uses rewards from groups of sampled responses to construct relative advantages. We used its multi-turn implementation in verl-agent as a teacher-free baseline (Shao et al., 2024).
- **GiGPO:** Group-in-Group Policy Optimization combines episode-level comparison with step-level grouping of actions from shared environment states, providing a second RL backbone with finer-grained credit assignment (Feng et al., 2025).
- **OPD:** On-policy distillation evaluates the teacher on student-generated contexts. Our OPD baseline added a teacher-KL term to the underlying RL objective and applied it to all valid response tokens, including those from successful trajectories. This served as the direct comparison for ECOD's supervision allocation (Agarwal et al., 2024).
- **SOD:** Step-wise On-policy Distillation adjusts the strength of distillation according to teacher–student divergence at each interaction step. We use our shared-framework implementation of its step-weighted teacher signal within the shared multi-turn GRPO training system (Zhong et al., 2026).
- **TCOD:** Temporal Curriculum On-Policy Distillation progressively expands the depth of student interaction. We implemented its front-to-back curriculum, denoted **TCOD-F2B**, by increasing the training rollout horizon while retaining the GRPO-plus-OPD objective (Wang et al., 2026).
- **ATOD:** Annealed Turn-Aware On-Policy Distillation combines an annealed balance between OPD and RL with turn-level disagreement–uncertainty reweighting. We imple-

mented these teacher-side mechanisms on the same multi-turn GRPO backbone (Tan et al., 2026).

The core comparison evaluated RL alone, RL with OPD, and RL with ECOD under both GRPO and GiGPO. The compatibility comparisons evaluated the specified SOD, TCOD-F2B, and ATOD mechanisms with and without ECOD in a shared multi-turn GRPO framework. In each pair, ECOD acted only on the teacher contribution; the environment-derived reward signal and the base method's curriculum or weighting rule were retained.

The same ECOD runs are compared against the RL-only and OPD baselines in the two table groups (Tables 1 and 2).

A.7 Training details

Student models and hardware. Students were initialized from Qwen2.5-1.5B-Instruct and Qwen2.5-7B-Instruct for experiments on ALFWorld and WebShop. Each run used a single node with eight NVIDIA H200 GPUs. Our implementation used verl-agent, built on verl (Sheng et al., 2025), with FSDP for distributed policy optimization and vLLM (Kwon et al., 2023) for rollout generation. The training schedule comprised 150 updates, each with 16 task conditions and eight rollouts per condition, yielding 128 trajectories per update. We used a learning rate of 10^{-6}, a training sampling temperature of 1.0, and one PPO epoch per update. Table A2 summarizes the Qwen2.5-1.5B-Instruct settings and benchmark-specific differences.

Teacher construction. Teachers were trained from Qwen2.5-1.5B-Instruct or Qwen2.5-7B-Instruct, matching the student's model size. For each benchmark, model size, and RL backbone, we froze the teacher obtained from the corresponding RL-only run after 150 updates. GRPO-based OPD, SOD, TCOD-F2B, ATOD, and their ECOD-assisted variants used the GRPO-trained teacher; GiGPO-based configurations used the GiGPO-trained teacher. Teacher-assisted students started from the common pretrained model, not from the teacher weights. The teacher generated eight rollouts for each entry in the frozen 150×16 training task-condition schedule. ECOD used their completion outcomes and successful trajectory lengths, together with student outcomes from the current batch, to construct its allocation weights. Teacher-assisted variants shared the teacher within each benchmark, model size, and RL backbone; paired comparisons reused the training and validation task schedules. Teacher precollection and online teacher scoring incurred additional cost beyond the student rollout budget.

Objectives and matching. For Qwen2.5-1.5B-Instruct, the RL-only, OPD, ECOD, and TCOD-F2B configurations retained a reference-policy KL coefficient of 0.01; those using an explicit teacher-KL term used a separate coefficient of 0.02. SOD and ATOD instead incorporated teacher feedback into their advantages, without an additional reference-KL or separate teacher-KL loss. Within these pairs, the teacher and non-ECOD hyperparameters were kept the same. Update and rollout counts are matched, while action and token consumption vary with TCOD-F2B's curriculum horizon.

7B WebShop teacher strengths. Experiments with Qwen2.5-7B-Instruct on WebShop used the same task configuration, update budget, and rollout group size, with a teacher of the same size trained using the corresponding RL backbone. ECOD used a teacher-KL coefficient of 0.08 for GRPO, GiGPO, and TCOD-F2B, a teacher-advantage coefficient of 4.0 for SOD, and four times the ATOD κ schedule. The corresponding baselines retained their original teacher-signal strengths.

Table A2. Training settings for the Qwen2.5-1.5B-Instruct experiments. Mini-batch sizes count response-turn samples, not complete trajectories.

Setting	ALFWorld	WebShop
Student initialization	Qwen2.5-1.5B-Instruct	Qwen2.5-1.5B-Instruct
Hardware	1 node × 8 NVIDIA H200	1 node × 8 NVIDIA H200
Training updates	150	150
Task conditions per update	16	16
Rollouts per condition	8	8
Trajectories per update	128	128
Learning rate	10^{-6}	10^{-6}

Setting	ALFWorld	WebShop
PPO epochs per update	1	1
PPO mini-batch size	256	64
PPO clip ratio	0.2	0.2
Entropy coefficient	0.001	0.001
Maximum action turns	50	15
History length	2	2
Maximum prompt length per turn	2,048 tokens	4,096 tokens
Maximum response length per turn	512 tokens	512 tokens
Training / evaluation temperature	1.0 / 0.4	1.0 / 0.4
Validation frequency	Every 5 updates	Every 5 updates
Validation batch	128	128
Invalid-action penalty	0.1	0.1
GiGPO discount γ	0.95	0.95
GiGPO step-advantage weight	1.0	1.0
GiGPO normalization	mean_std_norm	mean_norm

The maximum action horizon in Table A2 applies throughout evaluation and to non-curriculum training. For TCOD-F2B training, the horizon at update $m \geq 1$ was $K_m = \min(T_{\max}, 1 + \lfloor m/2 \rfloor)$, where $T_{\max}$ is the benchmark limit. Teacher statistics were collected at the full horizon.

Table A3. Method-specific settings for the 1.5B comparisons. The 7B WebShop teacher-strength adjustments are specified above.

Component	Setting
Explicit teacher-KL objectives	Reference KL coefficient 0.01; teacher KL coefficient 0.02
SOD	Divergence-ratio stabilization $\epsilon = 10^{-6}$; weight cap $1 + \delta$, $\delta = 0.2$
TCOD-F2B	Initial horizon 1; growth interval 2 updates; benchmark-specific cap
ATOD	OPD weight $\kappa : 1.0 \rightarrow 0.1$; RL weight $\rho : 1.0 \rightarrow 2.0$, over 80 updates; turn-level disagreement–uncertainty reweighting enabled
ECOD	Eight teacher rollouts per task condition; current-batch student outcomes; Beta(1,1) competence priors; empirical survival of successful teacher lengths

A.8 Evaluation details

Evaluation used the student alone, without teacher intervention, at a sampling temperature of 0.4 and the full interaction horizon of each benchmark. Our primary metric was **success rate**, the percentage of episodes satisfying the environment's task-completion criterion. On ALFWorld, we reported results separately for the full seen and unseen evaluation sets. On WebShop, we additionally reported **task score**, the mean normalized reward expressed as a percentage of its maximum value of 1, to quantify partial satisfaction of the shopping requirements.

Reported checkpoints. Main results use the designated checkpoint for each configuration; auxiliary metrics use the same checkpoint. Ablations use update 150.

Repeated evaluations kept the task sequence fixed and varied the decoding seed. For Qwen2.5-1.5B-Instruct, the core RL/OPD/ECOD comparison used eight decoding runs for each ALFWorld split. The SOD, TCOD-F2B, and ATOD comparisons used one run on ALFWorld seen and eight on unseen; all reported WebShop methods used eight runs on the same 128-task schedule. Repeated runs are summarized by their mean and sample standard deviation. Main tables omit the seen standard deviation: the six core 1.5B RL/OPD/ECOD configurations have eight identical seen scores,

whereas the remaining seen entries are single-run scores. The Qwen2.5-7B-Instruct experiments on ALFWorld used one seen run and eight unseen runs per evaluated model. All runs used training seed 0; the repetitions quantify decoding variability at fixed weights, not variability across independent training runs.

The abstract averages paired relative success-rate changes across SOD, TCOD-F2B, and ATOD with both student models, giving six equally weighted comparisons per benchmark. Each change is 100 times the ECOD–baseline difference divided by the baseline, computed from unrounded means and retaining negative values. ALFWorld uses unseen results; the two benchmarks are averaged separately.

A.9 Matched-allocation ablations

The WebShop component ablations used Qwen2.5-1.5B-Instruct with GRPO, eight H200 GPUs, and 150 training updates. They shared the initialization, frozen teacher, task schedules, and optimization settings of the full ECOD reference. The teacher-KL coefficient remained 0.02. GRPO, OPD, and full ECOD results were reused from the matched main-experiment protocol; the three factor-removal controls and shared-prior control were trained separately.

Let i index response-turn samples in a control's current rollout batch X, n_i denote the number of valid response tokens, and $g_i = C_i F_i H_{T,i}$ denote the full ECOD weight evaluated on that same batch. For a control weight b_i, we used

$$D(X) = \sum_i n_i g_i, \qquad \widetilde{b}_i = \begin{cases} b_i \dfrac{D(X)}{\sum_j n_j b_j}, & D(X) > 0, \\ 0, & D(X) = 0. \end{cases} \tag{13}$$

The unscaled controls were $b_i = F_i H_{T,i}$ (w/o competence weighting), $b_i = C_i H_{T,i}$ (w/o outcome selection), $b_i = C_i F_i$ (w/o temporal weighting), and $b_i = C_i F_i \overline{H}(t_i)$ (pooled temporal prior). The shared curve was computed by pooling all 11,244 successful teacher trajectories from the 2,400 training task-condition entries, with equal weight per trajectory:

$$\overline{H}(t) = \frac{1}{11244} \sum_{\ell=1}^{11244} \mathbf{1}[L_\ell > t]. \tag{14}$$

Matching protocol. The prior uses training teacher trajectories, and the loss denominator counts all valid response tokens. Each control matches the batch-level weighted-token budget of full ECOD on the same rollouts, while redistributing that budget according to the ablated allocation rule. Matching concerns this total, not realized KL, gradient magnitude, or token count.

Numerical checks. Across the completed runs, the maximum relative mismatch in token-weighted supervision totals was below 8×10^{-16}. A positive target with zero control support is treated as an error.

A.10 Teacher-strength and allocation controls

We used ALFWorld with Qwen2.5-1.5B-Instruct and GRPO, retaining the initialization, frozen teacher, training task sequence, and optimization settings of the corresponding main experiment. Every configuration was evaluated at the fixed training endpoint, step 150, on the same 134 unseen tasks with temperature 0.4 and no teacher intervention. Table A4 reports means and sample standard deviations over decoding seeds 0–7 for one training run per configuration. GRPO, OPD, and full ECOD reuse their existing endpoint evaluations.

Strength controls. OPD applies a uniform teacher coefficient of 0.02 or 0.002. Linear annealing uses $\lambda(s) = 0.004(150 - s)/149$ for updates $s = 1, \ldots, 150$; its update-average coefficient is 0.002, matching the smaller constant, without equalizing token counts or cumulative KL. All three coefficient schedules were prespecified and are reported. The comparison used mean Seen validation success over updates 130, 135, 140, 145, and 150. The constant 0.002 control ranked highest

(76.56%) and was selected, followed by linear annealing (76.41%) and the constant 0.02 control (76.25%).

Allocation controls. Uniform matching sets $b_i = 1$ in Eq. (13), giving every valid response token the same weight $D(X)/\sum_i n_i$, with base teacher coefficient 0.02. The target $D(X)$ is computed using full ECOD on the control's own current rollouts. The maximum relative matching error over 150 updates was 9.73×10^{-15}; 147 batches changed allocation and three had zero target budget. The additional control without competence weighting uses $b_i = F_i H_{T,i}$ under the same matching rule.

Frozen-budget control. Frozen-budget uniform allocation replays the weighted-token budget D_m^{full} recorded at each update m of the full ECOD run. Every valid response token in the control batch receives weight $D_m^{\text{full}}/N_m^{\text{ctrl}}$, where N_m^{ctrl} is its valid response-token count; the base teacher coefficient remains 0.02. All 150 update budgets are matched, with a maximum relative allocation error of 4.64×10^{-15} before actor-side floating-point conversion. This control preserves the reference budget sequence while replacing selective allocation with uniform weighting.

Table A4. ALFWorld teacher-strength and allocation controls (1.5B, GRPO). Fixed step 150; mean ± sample s.d. across eight decoding runs.

Configuration	Unseen success (%)
GRPO	57.56 ± 1.23
OPD, fixed 0.02	71.27 ± 2.22
OPD, fixed 0.002	72.20 ± 2.70
OPD, linear 0.004 to 0	67.35 ± 2.79
Current-batch uniform	64.93 ± 0.00
Frozen-budget uniform	72.29 ± 3.15
w/o competence weighting	71.46 ± 0.77
ECOD	**74.16 ± 0.97**

A.11 Learning-dynamics protocol

Figure 5 uses unsmoothed records from one training run per configuration. Shading covers updates 126–150; annotations report the relative change between the ECOD and baseline means over the five validation events in that interval. Teacher coefficients include each method's native weighting or annealing. SOD and ATOD use response-row means, whereas TCOD-F2B uses token-weighted means. Teacher coefficients are reported on method-specific scales and compared within each method pair.

A.12 Supervision coverage and component distributions

To characterize how ECOD allocates supervision beyond average teacher coefficients, we reconstruct its components from GRPO-based training rollouts and cached teacher records on ALFWorld and WebShop with Qwen2.5-1.5B-Instruct and Qwen2.5-7B-Instruct. Each setting includes one training run, all 150 updates, 2,400 task conditions, and 19,200 trajectories.

Measurement. Active-turn coverage is the fraction of recorded action turns assigned a positive ECOD weight. We weight task conditions equally for q, $\mu_T - \mu_S$, and C, trajectories equally for F, and active turns equally for H_T and g. All distributions retain zeros, except the explicitly conditional nonzero-weight summary.

Observed allocation. Across all updates, nonzero weights cover 41.70% and 24.40% of active turns on ALFWorld, and 30.66% and 21.35% on WebShop, for the 1.5B and 7B students, respectively (Figure A1). Figure A2 reports the corresponding component distributions. Figure A2 shows how the competence, outcome, and temporal factors evolve across training stages.

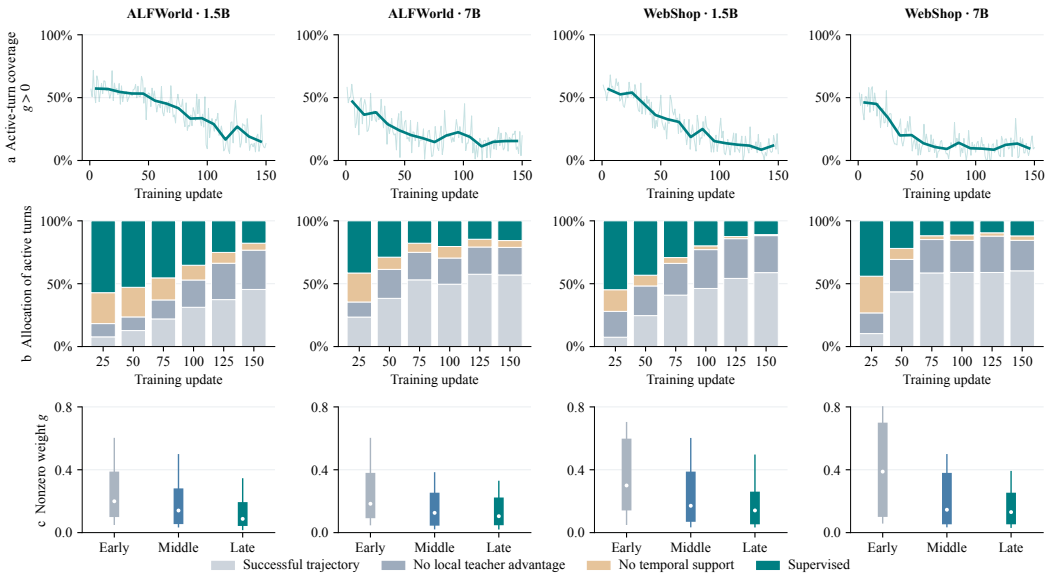

Figure A1: **Supervision allocation across training.** Rows show active-turn coverage, its ordered outcome/competence/temporal partition, and nonzero weights. Curves show raw updates and ten-update pools; bars pool 25 updates. Weight summaries show median, interquartile, and 10–90% ranges for early/middle/late stages (50 updates each).

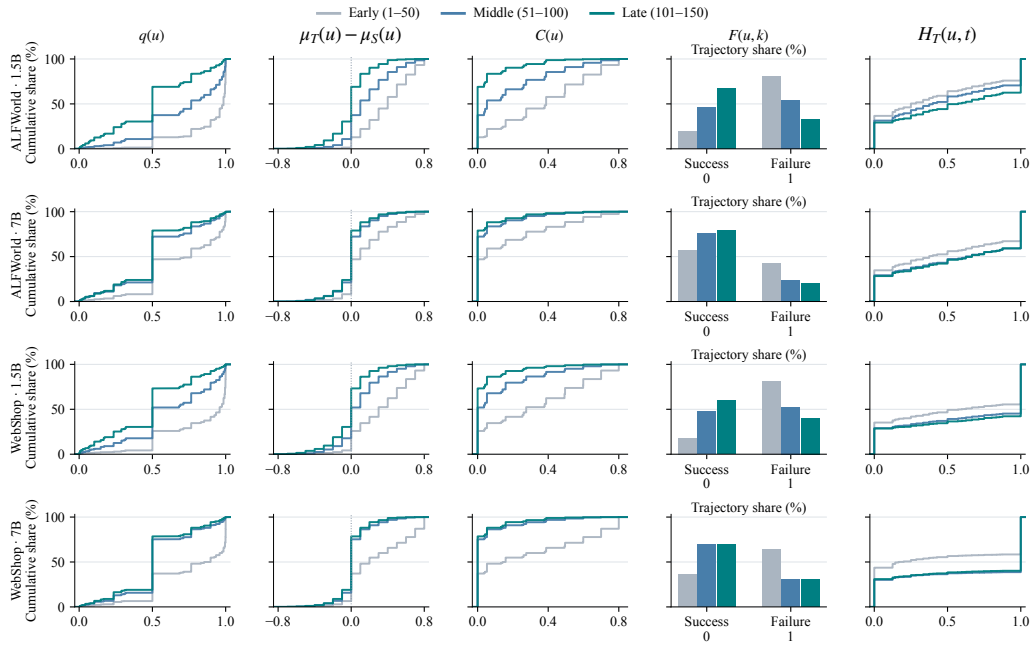

Figure A2: **Component distributions across training stages.** Curves show empirical cumulative distributions from one training run per setting; bars show the binary distribution of F. Gray/blue/teal denote updates 1–50/51–100/101–150. Each stage contains 800 task conditions and 6,400 trajectories per setting.