TEACHER–STUDENT GAPS ARE NOT ENOUGH: OUTCOME-GUIDED ON-POLICY DISTILLATION FOR MULTI-TURN AUTONOMOUS AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) trains a student on its own trajectories with dense teacher supervision. Recent work on OPD for multi-turn autonomous agents often treats large teacher–student token-level distributional gaps as promising intervention points, linking larger gaps to a greater need for correction. Yet, our empirical analysis reveals a **supervision–benefit mismatch**: large gaps can be benign, while small gaps can be outcome-critical. Teacher–student gaps capture differences at the current turn, whereas the benefit of teacher guidance depends on how the current student interacts with the environment afterward. The student may still succeed despite choosing an action that differs from the teacher's, while a teacher-preferred action may lead to a state from which the student cannot complete the task. Local gaps alone are therefore not enough to determine whether teacher guidance benefits the current student. Effective supervision should instead emphasize guidance that the current student can translate into better final task outcomes. Accordingly, we propose **Outcome-Guided On-Policy Distillation (OG-OPD)**, which applies trajectory-relative weighting to teacher supervision and calibrates these weights using final task outcomes from paired student continuations. This calibration selectively strengthens supervision on the student's original trajectories at turns where teacher guidance benefits the current student. Across ALFWorld, ScienceWorld, and WebShop, OG-OPD consistently outperforms baselines under diverse settings. It improves task success rates by 3.6–17.7 percentage points over vanilla OPD and by up to 7.0 percentage points over the strongest baseline.

1 INTRODUCTION

On-policy distillation (OPD) trains students on their own trajectories (Agarwal et al., 2024; Lin et al., 2020) with dense teacher supervision (Lu & Thinking Machines Lab, 2025; Fu et al., 2026). Learning at student-visited states helps reduce the distribution mismatch between training and in-

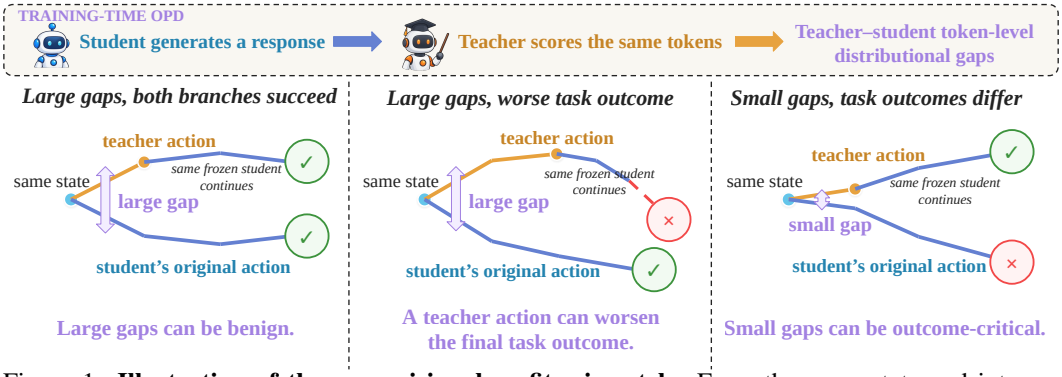


Figure 1: **Illustration of the supervision–benefit mismatch.** From the same state and interaction history, one branch takes the student's original action and the other a teacher-proposed action. Both branches then continue interacting under the same frozen student policy. Moving toward the teacher's preferred action does not necessarily improve the current student's final task outcome.

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.

ference (Bengio et al., 2015; Ranzato et al., 2016). At each turn, the teacher provides token-level probability feedback on the student's response (Jin et al., 2026; Ko et al., 2024; Jia et al., 2026), encouraging the student to narrow teacher–student token-level distributional gaps (Gu et al., 2024; Jang et al., 2026; Shao et al., 2026). However, local teacher–student gaps alone are not enough to determine whether teacher guidance benefits the current student. This limitation is especially important for multi-turn autonomous agents, where teacher supervision spans successive environment interactions (Wang et al., 2026; Liao et al., 2026). Actions at one turn shape subsequent observations (Yao et al., 2023) and the states in which later decisions are made (Ross et al., 2011). Errors can compound across successive turns (Ross & Bagnell, 2010; Li et al., 2026). Effective supervision must therefore consider both where to intervene and what the current student achieves afterward.

Against this background, existing agentic OPD methods organize teacher supervision across turns through rollout construction and teacher intervention. Some progressively extend student rollouts or shorten prefixes of successful teacher trajectories (Wang et al., 2026), giving the student control over more turns as training progresses. Others interleave teacher and student turns within a trajectory while gradually reducing the probability of teacher intervention (Li et al., 2026). For more targeted guidance, recent work uses teacher–student token-level distributional gaps to identify promising intervention points and steer subsequent student behavior toward the teacher (Chen et al., 2026). Yet, does teacher guidance at these turns actually improve the current student's final task outcome?

To examine this relationship, we conduct an empirical analysis on ScienceWorld (Wang et al., 2022) and WebShop (Yao et al., 2022). Our findings reveal a **supervision–benefit mismatch**: large gaps can be benign, while small gaps can be outcome-critical. A teacher-proposed action can change subsequent states and decisions, but its benefit depends on the current student's ability to complete the task from the resulting state. Assessing this benefit requires comparing the current student's final task outcomes after its original action and after the teacher action. Local gaps remain useful for locating potential intervention points, while this comparison provides evidence for deciding which turns warrant stronger teacher supervision. Figure 1 shows a schematic, with details in Section 3.1.

Accordingly, we propose **Outcome-Guided On-Policy Distillation (OG-OPD)**. First, OG-OPD uses trajectory-relative weighting to allocate teacher supervision and identifies candidate turns based on the current gap and an abrupt increase from the preceding turn. It then calibrates these weights using final task outcomes from paired student continuations. When the teacher action improves the current student's final task outcome, OG-OPD selectively strengthens supervision at the corresponding turn on the student's own trajectory. Otherwise, it retains the pre-calibration weight. Teacher responses and their continuations are used only for calibration, not as additional distillation targets.

We evaluate OG-OPD on ALFWorld (Shridhar et al., 2021), ScienceWorld, and WebShop under different teacher–student configurations. OG-OPD achieves the highest mean task success rate among the compared methods in every evaluated setting. On WebShop, for example, the Qwen3-1.7B student achieves a 45.7% success rate, compared with 28.0% for vanilla OPD and 38.7% for the strongest competing baseline. These gains extend to higher task scores on both ScienceWorld and WebShop. OG-OPD also requires fewer interaction turns than vanilla OPD in most comparisons.

Our main contributions are as follows. (1) **Supervision–benefit mismatch.** Moving toward the teacher does not necessarily benefit the current student. Across the two benchmarks, teacher harm averages 10.8% of large-gap decisions, exceeding teacher rescue at 8.8%. (2) **Outcome-Guided On-Policy Distillation.** OG-OPD builds on trajectory-relative weighting and uses task outcomes to decide where to strengthen teacher supervision. (3) **Broad validation.** Experiments on three benchmarks with different teacher–student settings show consistent gains in task success over baselines.

2 PRELIMINARIES

2.1 MULTI-TURN INTERACTION

At turn k, the frozen rollout student $p_\circ$ generates response z_k from history c_k. The action parser $\mathcal{A}$ maps z_k to action u_k, which changes environment state s_k according to transition law $\mathsf{T}_{\mathrm{env}}$:

$$\begin{aligned} z_k \sim p_\circ(\cdot \mid c_k), \qquad u_k = \mathcal{A}(z_k), \\ (s_{k+1}, o_{k+1}) \sim \mathsf{T}_{\mathrm{env}}(\cdot \mid s_k, u_k), \qquad c_{k+1} = c_k \oplus (z_k, o_{k+1}), \end{aligned} \tag{1}$$

Here s_k need not be fully observable and $\oplus$ appends the response and observation. Autoregressive generation gives $p_\circ(z_k \mid c_k) = \prod_j p_\circ(z_{k,j} \mid v_{k,j})$, with $v_{k,j} = (c_k, z_{k,<j})$. Trajectory ζ ends at termination or the interaction limit. Its final task outcome $\mathsf{S}(\zeta) \in \{0, 1\}$ indicates benchmark success. Turn sums cover $0 \le k < K_\zeta$, where K_ζ counts the interaction turns in trajectory ζ.

2.2 On-policy Distillation

OPD provides dense teacher supervision on the student's own trajectories (Agarwal et al., 2024). For fixed teacher q_ϕ and trainable student p_θ, reverse KL distillation (Gu et al., 2024) minimizes

$$\mathcal{J}_{\mathrm{rKL}}(\theta) = \mathbb{E}_{v \sim d_{p_\circ}}\left[D_{\mathrm{KL}}(p_\theta(\cdot \mid v) \,\|\, q_\phi(\cdot \mid v))\right],$$
$$D_{\mathrm{KL}}(p_\theta(\cdot \mid v) \,\|\, q_\phi(\cdot \mid v)) = \sum_{x \in \mathcal{X}} p_\theta(x \mid v) \log \frac{p_\theta(x \mid v)}{q_\phi(x \mid v)}, \tag{2}$$

where $\mathcal{X}$ is the vocabulary. The token context distribution $d_{p_\circ}$ is sampled under the frozen rollout student $p_\circ$ and held fixed while the trainable student p_θ is optimized within each update.

For each sampled token $z_{k,j}$, the teacher supervision and importance ratio are

$$\psi_{k,j} = \log \frac{q_\phi(z_{k,j} \mid v_{k,j})}{p_\circ(z_{k,j} \mid v_{k,j})}, \qquad \rho_{k,j}(\theta) = \frac{p_\theta(z_{k,j} \mid v_{k,j})}{p_\circ(z_{k,j} \mid v_{k,j})}. \tag{3}$$

At fixed v, the expected signed log probability ratio under $p_\circ$ equals the negative reverse KL:

$$\mathbb{E}_{x \sim p_\circ(\cdot \mid v)}\left[\log \frac{q_\phi(x \mid v)}{p_\circ(x \mid v)}\right] = -D_{\mathrm{KL}}(p_\circ(\cdot \mid v) \,\|\, q_\phi(\cdot \mid v)). \tag{4}$$

Positive $\psi_{k,j}$ favors a higher sampled token probability, while negative values favor a lower one.

Let $\mathcal{M}_k(\zeta)$ index generated tokens with valid teacher scores, excluding prompts, observations, and padding, and let $Z = \sum_{\zeta \in \mathcal{D}} \sum_k |\mathcal{M}_k(\zeta)|$ count them in batch $\mathcal{D}$. The clipped policy surrogate (Schulman et al., 2017), defined in Appendix A, gives the turn-level and batch losses to minimize:

$$\mathcal{Q}_k(\theta; \zeta) = \sum_{j \in \mathcal{M}_k(\zeta)} \ell_{\mathrm{pol}}(\rho_{k,j}(\theta), \mathrm{sg}[\psi_{k,j}]),$$
$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \frac{1}{Z} \sum_{\zeta \in \mathcal{D}} \sum_k \mathcal{Q}_k(\theta; \zeta). \tag{5}$$

Here $\mathrm{sg}[\cdot]$ holds $\psi_{k,j}$ fixed during differentiation. OG-OPD changes only the turn-level weights.

We quantify the magnitude of teacher–student token-level distributional gaps at each turn using

$$\chi_k = \frac{1}{|\mathcal{M}_k(\zeta)|} \sum_{j \in \mathcal{M}_k(\zeta)} |\psi_{k,j}|, \qquad k \in \mathcal{V}(\zeta). \tag{6}$$

Here $\mathcal{V}(\zeta) = \{k : |\mathcal{M}_k(\zeta)| > 0\}$ retains original turn indices. The sampled gap magnitude χ_k is distinct from signed token-level update directions and full-vocabulary KL divergence.

3 Outcome-Guided On-Policy Distillation

We first present an empirical analysis of the supervision–benefit mismatch. Motivated by these findings, we propose OG-OPD, building on trajectory-relative weighting and candidate selection. Its core component, outcome-based calibration, uses final task outcomes from paired student continuations to selectively strengthen teacher supervision on the student's own trajectories.

3.1 Empirical Analysis

We analyze paired student continuations on ScienceWorld and WebShop using a fixed Qwen3-32B teacher and Qwen3-1.7B student. At sampled student-visited states, one branch takes the student's original action and the other the teacher action, with both starting from the same state and history and continuing under the same frozen student policy. Within each benchmark, the highest and lowest of five quantile groups define the large-gap and small-gap groups. Each decision is classified from five matched trials. Appendix B describes state sampling and the full outcome classification rules.

Figure 2 summarizes final task outcomes across gap groups. In the large-gap group, benign gaps account for an average of 5.4% of decisions, where both the student's original action and the teacher action lead to success. Teacher rescue and teacher harm occur in 8.8% and 10.8% of decisions, respectively. These findings indicate that a large gap can be benign, while the teacher action can either improve or worsen the final task outcome. In the small-gap group, 7.6% of decisions are outcome-critical, with the two branches leading to different final task outcomes. Together, these results reveal the supervision–benefit mismatch. Beyond these averages, the relative frequencies of teacher rescue and teacher harm vary across benchmarks. Within the large-gap group, teacher harm occurs more often than teacher rescue on WebShop, whereas the reverse holds on ScienceWorld. Appendix B.1 further examines this pattern under gap magnitude ranking.

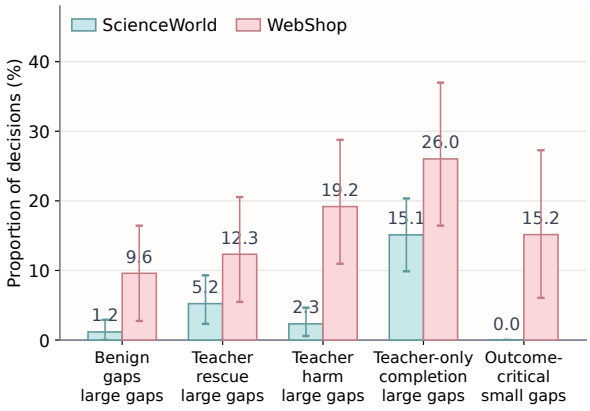


Figure 2: **Empirical analysis of final task outcomes.** Bars show decision proportions by outcome category within each benchmark's gap groups. Error bars indicate 95% decision-level bootstrap confidence intervals.

Beyond gap magnitude, we examine the role of the continuation policy by comparing teacher and student continuations after the same teacher action. In an average of 20.6% of large-gap decisions across the two benchmarks, the teacher frequently completes the task, while the current student rarely succeeds from the same resulting state. We refer to these cases as teacher-only completion. This suggests that the value of teacher guidance depends on the current student's ability to continue from the resulting state and complete the remaining task successfully. These findings motivate OG-OPD to use paired student continuations when deciding where to strengthen teacher supervision.

3.2 Trajectory-Relative Weighting and Candidate Selection

Figure 3 illustrates trajectory-relative weighting and candidate selection. For the gap χ_k in Equation 6, we define its logarithmic magnitude ν_k and change from the preceding turn υ_k as

$$\nu_k = \log(\varepsilon_0 + \chi_k), \qquad \upsilon_k = \nu_k - \nu_{k-1} = \log \frac{\varepsilon_0 + \chi_k}{\varepsilon_0 + \chi_{k-1}}, \tag{7}$$

where $\varepsilon_0 > 0$ stabilizes the logarithm. The change υ_k requires adjacent valid turns $k, k-1 \in \mathcal{V}(\zeta)$.

Since χ_k averages $|\psi_{k,j}|$, larger gaps imply larger supervision coefficients on average. Trajectory-relative weighting moderates this variation using the first valid turn $k_0 = \min \mathcal{V}(\zeta)$ as a reference. Each valid turn in the student's own trajectory receives the pre-calibration weight

$$\beta_k = \begin{cases} 1, & k = k_0, \\ \min\{\beta_{\max}, \exp(\nu_{k_0} - \nu_k)\}, & k > k_0, \end{cases} \qquad \beta_{\max} \geq 1. \tag{8}$$

Gaps above the reference reduce the weight, while gaps below it increase the weight up to $\beta_{\max}$. Using Equation 6, the average magnitude of the weighted token-level teacher supervision is

$$\frac{1}{|\mathcal{M}_k(\zeta)|} \sum_{j \in \mathcal{M}_k(\zeta)} |\beta_k \psi_{k,j}| = \beta_k \chi_k = \min\left\{ \beta_{\max} \chi_k, \frac{(\varepsilon_0 + \chi_{k_0})\chi_k}{\varepsilon_0 + \chi_k} \right\}. \tag{9}$$

With ε_0 negligible relative to the gaps and the cap inactive, $\beta_k \chi_k \approx \chi_{k_0}$. This sets a common supervision scale across valid turns in the trajectory, but does not determine whether teacher guidance benefits the current student in terms of its final task outcome under its continuation policy.

To limit additional environment execution, we evaluate paired student continuations only at selected turns. Candidates require a sufficiently large gap and, after the initial turn, a positive increase meeting its threshold. Final task outcomes determine whether to strengthen supervision. Let $\mathcal{H}(\zeta)$ contain turns with valid scores and replay information. We define the candidate set as

$$\mathcal{K}(\zeta) = \left\{ k \in \mathcal{H}(\zeta) \;\middle|\; \nu_k \geq \vartheta_\nu \;\wedge\; \left[k = 0 \;\vee\; (\upsilon_k > 0 \;\wedge\; \upsilon_k \geq \vartheta_\upsilon) \right] \right\}. \tag{10}$$

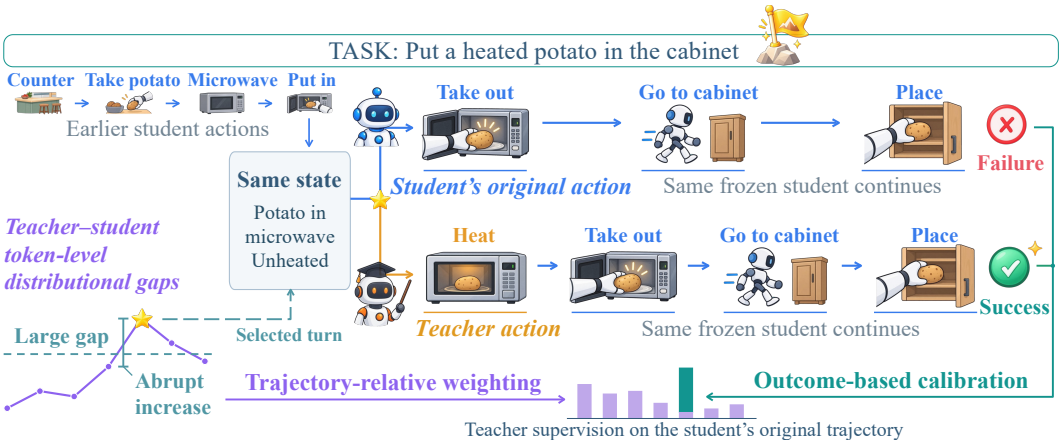


Figure 3: **Overview of Outcome-Guided On-Policy Distillation.** Trajectory-relative weighting assigns supervision weights, while gap magnitude and abrupt increases identify candidate turns. Outcome-based calibration compares paired student continuations and strengthens supervision on the student's original trajectory when teacher guidance improves the final task outcome.

Thresholds are batch quantiles with equal total weight per contributing trajectory. Only positive valid changes enter ϑ_v. The magnitude-only exception applies to $k = 0 \in \mathcal{H}(\zeta)$, not a later first valid turn $k_0 > 0$. Within a bounded proposal budget, we check candidates chronologically and select the first replay-valid turn $k^\star$ whose teacher response yields a legal action different from the student's original action. Only this turn undergoes paired student continuations for outcome-based calibration in Section 3.3. Appendix C gives further details of this execution procedure.

3.3 Outcome-Based Calibration

Outcome-based calibration increases the selected candidate turn's pre-calibration weight only when the current student fails with its original action but succeeds with the teacher action.

Let $z^S_{k^\star} = z_{k^\star}$ denote the student's original response and $z^T_{k^\star}$ the selected teacher response sampled from $q_\phi(\cdot \mid c_{k^\star})$. We restore state $s_{k^\star}$ and history $c_{k^\star}$ from the student's own trajectory. Define $\mathcal{U}(s, c, z; p_\circ, \varsigma)$ to execute $\mathcal{A}(z)$, retain z in history, and continue under the frozen current student $p_\circ$ with seed schedule ς. The paired student continuations are then generated as follows

$$\begin{aligned} \tilde{\zeta}^S &= \mathcal{U}(s_{k^\star}, c_{k^\star}, z^S_{k^\star}; p_\circ, \varsigma), \\ \tilde{\zeta}^T &= \mathcal{U}(s_{k^\star}, c_{k^\star}, z^T_{k^\star}; p_\circ, \varsigma). \end{aligned} \tag{11}$$

Superscript T denotes the teacher response, not teacher continuation. Both use the same frozen current student, initial state, history, decoding settings, remaining interaction budget, and seed schedule. Responses may change subsequent states, so we test whether the current student can complete the task after taking the teacher action, rather than whether the teacher can. Training uses one pair at the selected turn, whereas the empirical analysis repeats matched trials per sampled decision.

Let $\mathcal{D}_{\text{pair}} \subseteq \mathcal{D}$ contain trajectories with valid completed paired student continuations. Define

$$\Delta_\zeta = \mathsf{S}(\tilde{\zeta}^T) - \mathsf{S}(\tilde{\zeta}^S) \in \{-1, 0, 1\}. \tag{12}$$

Here $\Delta_\zeta = 1$ denotes failure under the original action and success under the teacher action. $\Delta_\zeta = -1$ denotes the reverse, and $\Delta_\zeta = 0$ denotes no change in final task outcome. Only the positive case permits additional distillation weight at the selected candidate turn.

$$\mathsf{g}_\zeta = \mathsf{S}(\tilde{\zeta}^T)\big(1 - \mathsf{S}(\tilde{\zeta}^S)\big) = \mathbb{I}[\Delta_\zeta = 1] = [\Delta_\zeta]_+ . \tag{13}$$

Success with the teacher action provides positive execution evidence only when the original action fails, excluding cases where the current student succeeds under both responses.

The calibrated weight ω_k adds the outcome-based increment γ_ζ to the pre-calibration weight at the selected turn. For weight floor $\beta_\uparrow > 0$ and $[x]_+ = \max\{x, 0\}$, the resulting weights are

$$\gamma_\zeta = \mathsf{g}_\zeta \, [\beta_\uparrow - \beta_{k^\star}]_+ , \qquad \omega_k = \begin{cases} \beta_k + \gamma_\zeta, & k = k^\star, \\ \beta_k, & k \neq k^\star. \end{cases} \tag{14}$$

When $\mathrm{g}_\zeta = 1$, the calibrated weight is $\max\{\beta_{k^\star}, \beta_\uparrow\}$. Paired outcomes determine whether to add supervision, while the floor and pre-calibration weight determine the increase. Positive execution evidence can thus strengthen supervision at a previously downweighted turn. Otherwise, or without valid paired student continuations, $\gamma_\zeta = 0$ and pre-calibration weights remain unchanged. Since $\beta_{\max}$ is not reapplied, a higher floor can raise the calibrated weight above the pre-calibration cap.

For a fixed batch $\mathcal{D}$ of the student's own trajectories, calibrated weights scale the original response losses $\mathcal{Q}_k(\theta;\zeta)$ from Section 2.2. Teacher responses and paired student continuations are used only for outcome-based calibration, not as additional distillation targets. Keeping the original token normalization Z unchanged, we write the resulting weighted training objective as

$$\begin{aligned}\mathcal{L}_{\mathrm{OG-OPD}}(\theta) &= \frac{1}{Z}\sum_{\zeta\in\mathcal{D}}\sum_{k\in\mathcal{V}(\zeta)} \mathrm{sg}[\omega_k]\,\mathcal{Q}_k(\theta;\zeta) \\ &= \underbrace{\frac{1}{Z}\sum_{\zeta\in\mathcal{D}}\sum_{k\in\mathcal{V}(\zeta)} \mathrm{sg}[\beta_k]\,\mathcal{Q}_k(\theta;\zeta)}_{\mathcal{L}_{\mathrm{base}}(\theta)} + \frac{1}{Z}\sum_{\zeta\in\mathcal{D}_{\mathrm{pair}}} \mathrm{sg}[\gamma_\zeta]\,\mathcal{Q}_{k^\star}(\theta;\zeta).\end{aligned} \tag{15}$$

The first term retains dense teacher supervision across valid turns using their pre-calibration weights. The second supplements trajectory-relative weighting with supervision on the selected turn's original student response only when $\gamma_\zeta > 0$, leaving other turns' pre-calibration weights unchanged.

For this fixed batch, holding recorded $\psi_{k,j}$ and weights fixed during differentiation gives

$$\ell_{\mathrm{pol}}(\rho_{k,j}(\theta), \mathrm{sg}[\omega_k\psi_{k,j}]) = \mathrm{sg}[\omega_k]\,\ell_{\mathrm{pol}}(\rho_{k,j}(\theta), \mathrm{sg}[\psi_{k,j}]), \qquad \omega_k \geq 0. \tag{16}$$

Thus, turn weighting scales token-level teacher supervision while preserving the directions defined in Section 2.2. With the original token mask $\mathcal{M}_k(\zeta)$ and normalization Z unchanged, outcome-based calibration adds a gradient contribution only from the selected candidate turns

$$\nabla_\theta\mathcal{L}_{\mathrm{OG-OPD}}(\theta) - \nabla_\theta\mathcal{L}_{\mathrm{base}}(\theta) = \frac{1}{Z}\sum_{\zeta\in\mathcal{D}_{\mathrm{pair}}} \mathrm{sg}[\gamma_\zeta]\,\nabla_\theta\mathcal{Q}_{k^\star}(\theta;\zeta). \tag{17}$$

Paired outcomes provide evidence for strengthening supervision at the selected turn, while signed token-level teacher feedback determines the update directions on the original student response. Since Z counts valid original tokens rather than weights, other turns' loss coefficients remain unchanged. Appendix A provides the derivation and relates observed final task outcomes to training updates.

4 Experiments

4.1 Experimental Setup

Benchmarks and Models. We evaluate OG-OPD on ALFWorld, ScienceWorld, and WebShop, following TCOD (Wang et al., 2026) for environment configurations, prompt templates, and ALFWorld evaluation collections. All experiments use Qwen3 models (Yang et al., 2025). A Qwen3-32B teacher is paired with a Qwen3-1.7B student on all benchmarks, while ALFWorld and WebShop additionally use a Qwen3-8B-RL teacher with a Qwen3-4B student. Each RL teacher is trained with GiGPO (Feng et al., 2025) on its benchmark. We report success rate (SR), task score (Score), and mean interaction rounds (Rounds). Appendices D.1–D.2 detail teacher training and evaluation.

Training Setup. All methods are evaluated at training step 200, with results in the main, ablation, and efficiency tables averaged over three training seeds. Within each benchmark and model setting, all methods share the same training tasks, student initialization, teacher checkpoints, interaction protocol, and evaluation protocol, while retaining their original curriculum and prefix designs. Additional execution details, including action parsing and state replay, are provided in Appendix C.

Baselines. We compare OG-OPD with five agentic distillation baselines. Vanilla OPD provides dense teacher supervision on the student's own trajectories (Agarwal et al., 2024). TCOD-F2B progressively increases student rollout depth, whereas TCOD-B2F shortens successful teacher prefixes (Wang et al., 2026). Guided-OPD interleaves teacher and student turns while gradually increasing student control (Li et al., 2026). FutureBridge-OPD identifies large-gap positions, evaluates teacher guidance using teacher preference over student continuations, and adds accepted teacher responses as distillation targets (Chen et al., 2026). Zero-shot teachers and students serve as references.

Table 1: **Main results with the Qwen3-32B teacher and Qwen3-1.7B student.** Trained methods report mean ± standard deviation across three training seeds. SR (%) denotes success rate, with ALFWorld Overall pooling all splits. Score uses a 0–100 scale, and Rounds averages interactions per task. Dark/light green mark the best/second-best means among methods, with the best in bold.

	ALFWorld					ScienceWorld			WebShop		
Method	Seen SR ↑	Unseen SR ↑	Hard SR ↑	Overall SR ↑	Rounds ↓	SR ↑	Score ↑	Rounds ↓	SR ↑	Score ↑	Rounds ↓
Qwen3-32B teacher → Qwen3-1.7B student											
Student (zero-shot)	7.1	8.2	0.0	5.3	19.8	0.2	3.8	17.0	27.0	24.2	8.1
Teacher (zero-shot)	37.1	35.1	9.9	28.1	18.5	28.2	53.9	12.0	48.0	47.3	4.9
Vanilla OPD	26.7±3.0	27.6±3.3	4.1±1.4	20.1±2.1	**17.7±0.4**	13.0±1.4	48.1±1.2	12.6±0.3	28.0±2.6	27.5±2.6	6.8±0.2
TCOD-F2B	23.8±3.2	27.9±3.0	**7.4±0.8**	20.2±2.0	19.2±0.3	13.7±0.7	50.0±0.5	12.5±0.2	30.7±1.5	25.9±1.8	7.7±0.2
TCOD-B2F	31.0±3.6	**33.6±3.3**	5.8±1.4	24.1±2.3	18.5±0.3	14.3±0.5	48.3±0.6	11.9±0.2	33.7±3.1	33.5±3.4	6.1±0.3
Guided-OPD	26.7±5.1	28.4±4.9	5.8±1.7	20.8±3.3	18.4±0.5	15.8±0.8	45.2±2.0	12.4±0.4	34.3±9.6	30.8±8.9	5.9±0.4
FutureBridge-OPD	26.7±3.3	29.4±3.0	6.6±0.8	21.4±2.1	18.1±0.3	12.8±0.7	42.6±0.5	11.7±0.2	38.7±1.5	38.1±1.6	5.7±0.2
OG-OPD	**31.7±2.1**	32.3±2.3	6.6±0.8	**24.2±1.5**	18.0±0.2	**16.6±0.6**	**52.0±0.7**	**11.6±0.2**	**45.7±1.2**	**44.3±1.5**	**5.6±0.2**

Table 2: **Main results with the Qwen3-8B-RL teacher and Qwen3-4B student.** Each teacher is trained with GiGPO on its benchmark. Results for trained methods are mean ± standard deviation across three training seeds. SR (%), Score (0–100), and Rounds follow Table 1. Dark/light green mark the best/second-best means among compared methods, with the best in bold.

	ALFWorld					WebShop		
Method	Seen SR ↑	Unseen SR ↑	Hard SR ↑	Overall SR ↑	Rounds ↓	SR ↑	Score ↑	Rounds ↓
Qwen3-8B-RL teacher → Qwen3-4B student								
Student (zero-shot)	32.1	26.9	9.9	23.5	17.5	27.0	26.5	8.0
Teacher (zero-shot)	81.4	79.1	33.9	66.1	10.3	54.0	56.4	4.4
Vanilla OPD	71.4±3.7	67.4±3.5	20.9±2.4	54.6±2.7	11.3±0.3	51.7±1.5	55.2±2.2	4.2±0.1
TCOD-F2B	77.9±1.4	71.9±1.7	28.1±1.4	60.6±1.2	11.5±0.2	55.3±2.3	57.8±2.8	4.0±0.1
TCOD-B2F	77.1±2.1	75.6±1.9	**31.4±1.4**	62.6±1.6	11.3±0.2	54.7±1.2	56.6±1.5	4.1±0.1
Guided-OPD	74.3±3.8	73.4±3.5	19.3±2.5	57.1±2.8	11.2±0.3	52.3±4.6	53.7±3.6	4.3±0.2
FutureBridge-OPD	78.6±2.9	**77.1±2.8**	28.4±1.9	62.7±2.1	11.5±0.2	54.7±0.6	56.9±1.1	4.1±0.1
OG-OPD	**82.1±1.9**	76.4±1.9	30.0±1.3	**64.2±1.4**	**11.1±0.2**	**56.3±1.2**	**57.9±1.3**	**3.9±0.1**

4.2 Main Results and Training Dynamics

OG-OPD achieves the highest SR among the compared methods in every evaluated setting, with Overall SR reported for ALFWorld, as shown in Tables 1 and 2. Relative to vanilla OPD, the gains range from 3.6 to 17.7 percentage points. With the Qwen3-32B teacher and Qwen3-1.7B student, OG-OPD reaches 45.7% SR on WebShop, compared with 38.7% for FutureBridge-OPD, the strongest competing baseline. The advantage also holds with the task-trained Qwen3-8B-RL teacher and Qwen3-4B student. On ALFWorld, OG-OPD achieves 64.2% Overall SR, surpassing FutureBridge-OPD at 62.7%. Both students also outperform vanilla OPD on Seen, Unseen, and Hard tasks, showing consistent gains in unfamiliar environments and on difficult tasks.

With the Qwen3-32B teacher and Qwen3-1.7B student, OG-OPD improves SR over vanilla OPD by 17.7 percentage points on WebShop, compared with 3.6 points on ScienceWorld. This contrast is consistent with the empirical analysis in Section 3.1, which shows that teacher harm is more frequent than teacher rescue within WebShop's large-gap group, while the reverse holds on ScienceWorld. The larger improvement on WebShop, together with this empirical finding, supports the motivation for checking whether teacher guidance benefits the current student before strengthening supervision.

The gains extend beyond the binary success criterion used for outcome-based calibration. Across all evaluated ScienceWorld and WebShop settings, OG-OPD achieves the highest Score and the lowest Rounds among the compared methods. For Qwen3-1.7B, ScienceWorld Score increases from 48.1 with vanilla OPD to 52.0, while Rounds decreases from 12.6 to 11.6. On WebShop, Score increases

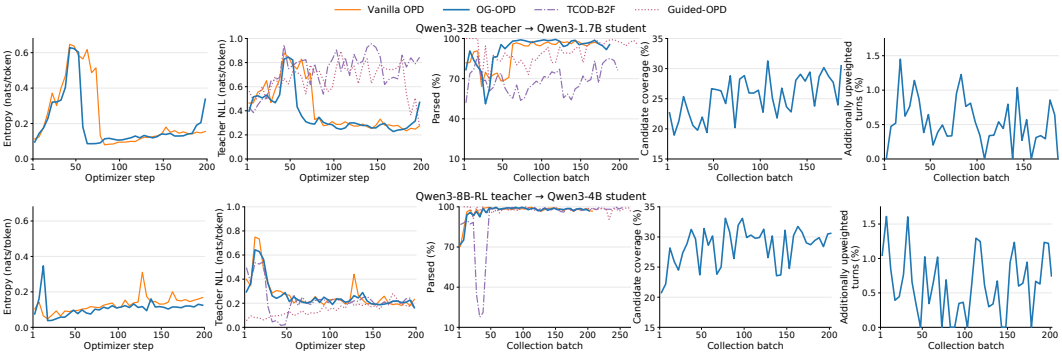


Figure 4: **Training dynamics on WebShop.** The top row shows the Qwen3-32B teacher with the Qwen3-1.7B student. The bottom row shows the Qwen3-8B-RL teacher with the Qwen3-4B student.

Table 3: **Ablation results.** The first group tests outcome-based calibration by removing it or directly upweighting candidate turns without verification. The second retains calibration and teacher response evaluation but replaces trajectory-relative weighting with uniform unit weights or gap-based candidate selection with random turn sampling. Trained methods report mean ± standard deviation across three training seeds. Bold marks each column's best mean among trained methods.

	ScienceWorld			WebShop		
Method	SR ↑	Score ↑	Rounds ↓	SR ↑	Score ↑	Rounds ↓
Qwen3-32B teacher → Qwen3-1.7B student						
Student (zero-shot)	0.2	3.8	17.0	27.0	24.2	8.1
Teacher (zero-shot)	28.2	53.9	12.0	48.0	47.3	4.9
Vanilla OPD	13.0±1.4	48.1±1.2	12.6±0.3	28.0±2.6	27.5±2.6	6.8±0.2
Outcome-Based Calibration						
w/o outcome-based calibration	3.8±1.1	38.5±1.9	12.5±0.3	32.0±2.6	30.5±2.5	6.4±0.2
Upweighting without verification	14.9±0.9	49.9±1.1	12.3±0.2	35.0±2.0	34.5±2.1	6.1±0.2
Weighting and Candidate Selection						
w/o trajectory-relative weighting	5.5±1.3	38.1±2.1	12.9±0.3	33.3±2.5	31.9±2.8	6.1±0.2
Random turn sampling	15.1±0.8	51.5±0.9	12.1±0.2	37.7±1.5	35.6±2.0	5.9±0.2
OG-OPD	**16.6**±0.6	**52.0**±0.7	**11.6**±0.2	**45.7**±1.2	**44.3**±1.5	**5.6**±0.2

from 27.5 to 44.3, while Rounds decreases from 6.8 to 5.6. The Qwen3-4B student on WebShop shows the same pattern. Higher task success accompanies higher task scores and shorter interactions, even though neither Score nor Rounds is used to trigger additional upweighting.

On WebShop, OG-OPD shows earlier improvements in teacher likelihood and action parseability than vanilla OPD. With the Qwen3-32B teacher and Qwen3-1.7B student, Teacher NLL falls earlier and Parsed recovers sooner, as shown in Figure 4. These improvements accompany selective additional upweighting. Across both teacher–student configurations, candidate turns account for roughly one quarter of eligible turns, whereas fewer than 1% of recorded turns receive additional weight when pooled over training. Thus, earlier response-level improvements coexist with additional upweighting of only a small subset of turns, while dense teacher supervision is retained across the student's own trajectories. These gains come at a per-step training cost of 1.42× vanilla OPD on ScienceWorld and 1.48× on WebShop with the Qwen3-1.7B student, representing a practical tradeoff between training cost and task performance. Appendix E provides training time details.

4.3 Ablation Study

We evaluate outcome-based calibration, trajectory-relative weighting, and candidate selection using the Qwen3-32B teacher and Qwen3-1.7B student, as shown in Table 3. We first test whether simpler weighting strategies can replace outcome-based calibration. The *w/o outcome-based calibration* variant retains the trajectory-relative weights without further adjustment, while *Upweighting without verification* directly increases the weight of the earliest gap-based candidate turn. Neither variant generates teacher responses or performs paired student continuations. Direct upweighting improves SR from 3.8% to 14.9% on ScienceWorld and from 32.0% to 35.0% on WebShop, but still falls short

of OG-OPD at 16.6% and 45.7%, respectively. Increasing candidate turn weights improves SR on both benchmarks, but does not match the gains from the full outcome-based calibration procedure.

With outcome-based calibration retained, we next examine the role of trajectory-relative weighting. The *w/o trajectory-relative weighting* variant replaces the pre-calibration weights with uniform unit weights while keeping candidate selection, teacher response evaluation, and calibration unchanged. Relative to OG-OPD, SR drops by 11.1 percentage points on ScienceWorld and 12.4 points on WebShop. These results show that trajectory-relative weighting and outcome-based calibration are complementary, as removing either component substantially reduces performance.

Finally, we examine the effect of candidate turn selection. The *Random turn sampling* variant randomly selects eligible turns without replacement from trajectories containing at least one candidate turn, with teacher response evaluation and outcome-based calibration unchanged. Relative to OG-OPD, SR drops by 1.5 percentage points on ScienceWorld and 8.0 points on WebShop. Gap-based candidate selection thus helps outcome-based calibration focus on more informative turns.

5 Related Work

On-Policy Distillation for Multi-Turn Agents. Learning from student-generated responses addresses the training–inference distribution mismatch in autoregressive distillation (Agarwal et al., 2024). Subsequent research spans objective design and rollout construction. Objective-level studies investigate reverse KL distillation (Gu et al., 2024), adaptive intermediate distributions (Jang et al., 2026), and entropy-dependent objectives (Jin et al., 2026). Complementary work combines skew KL objectives with adaptive reuse of student-generated outputs (Ko et al., 2024). In multi-turn environments, temporal curricula regulate student rollout depth (Wang et al., 2026), while turn-level guidance schedules teacher participation within a rollout (Li et al., 2026). Prefix replay organizes student continuations using previously collected teacher trajectories (Liao et al., 2026). These approaches shape the states visited during training and the distributions used as learning targets.

Selective Supervision and Reweighting. In chain-of-thought distillation, learned token weights emphasize key reasoning tokens (Feng et al., 2024). For OPD, supervision strength is adjusted across tokens or steps using entropy and teacher–student token-level distributional gaps (Xu et al., 2026), or final task outcomes and perplexity (Zheng et al., 2026). A trajectory-relative weighting rule interprets these gaps as a reliability signal and attenuates supervision as their magnitude increases relative to a reference (Zhong et al., 2026). Although its overall objective includes trajectory rewards, the distillation weights depend on relative gap magnitudes rather than the task outcomes induced by teacher actions. The weighting rule therefore does not distinguish downweighted turns where teacher guidance would help the current student from those where it would not.

Downstream Evaluation and Teacher Guidance. In reasoning tasks, sampled continuations provide automatic step-level supervision (Wang et al., 2024) and Monte Carlo estimates for credit assignment (Kazemnejad et al., 2025). Policy optimization also uses discounted future KL to shape token-level advantages (Ma et al., 2026) and repeated-state action groups for step-level credit assignment (Feng et al., 2025). In agent distillation, Chen et al. (2026) evaluate teacher responses through increases in the proportion of teacher-preferred tokens in paired student continuations, retaining accepted teacher responses as additional distillation targets. OG-OPD likewise evaluates teacher guidance through paired student continuations, but differs in both the validation signal and how it enters training. It uses final task outcomes from paired student continuations to decide whether to strengthen teacher supervision at the selected turn of the student's original trajectory, rather than adding accepted teacher responses as additional distillation targets for student policy optimization.

6 Conclusion

For multi-turn autonomous agents, teacher–student token-level distributional gaps are not enough to determine whether teacher guidance benefits the current student. Outcome-Guided On-Policy Distillation (OG-OPD) addresses this supervision–benefit mismatch by calibrating trajectory-relative weights with final task outcomes from paired student continuations while training on the student's original trajectories. Across three benchmarks and two teacher–student pairings, OG-OPD achieves the highest mean task success rate among the compared methods in every evaluated setting.

AI USE STATEMENT

Generative AI tools were used to polish the paper and assist with experimental code development. They were not used for any other tasks requiring disclosure. All AI-assisted work was reviewed by the authors. AI-assisted code was reviewed and tested for correctness, and language edits were checked to ensure they preserved the intended technical content. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

We evaluate language agents in benchmark environments, without live purchases or physical deployments. Improved benchmark performance does not establish safety in real-world settings, where erroneous actions may cause consequences not captured by these environments. Deployment therefore requires separate safety evaluations and appropriate safeguards. All datasets, environments, and models remain subject to their respective licenses and terms of use.

REPRODUCIBILITY STATEMENT

We provide the source code in the supplementary material accompanying this submission. Section 3 presents the empirical analysis and the OG-OPD algorithm, with optimization details, diagnostic protocols, and implementation procedures in Appendices A–C. Appendix D.1 specifies the model configurations, and Appendix D.2 describes the evaluation metrics and aggregation across training seeds. The training cost measurement protocol is provided in Appendix E. Together, these materials document the method and experimental procedures needed to reproduce the reported results.

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*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5be69a584901a26c521c2b51e40a4c20-Abstract-Conference.html.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/e995f98d56967d946471af29d7bf99f1-Abstract.html.

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.

Kaituo Feng, Changsheng Li, Xiaolu Zhang, Jun Zhou, Ye Yuan, and Guoren Wang. Keypoint-based progressive chain-of-thought distillation for LLMs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 13241–13255, 2024. URL https://proceedings.mlr.press/v235/feng24e.html.

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.

Zixuan Fu, Bingxiang He, Yuxin Zuo, Haohuan Huang, Jinqian Zhang, Ruhang Xiao, Cheng Qian, Qinyu Luo, Huan-ang Gao, Yudong Wang, Zhiyuan Liu, Ning Ding, and Chaojun Xiao. Rethinking on-policy distillation of large language models II: One training example. *arXiv preprint arXiv:2609.04172*, 2026. URL https://arxiv.org/abs/2609.04172.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5h0qf7IBZZ.

Ijun Jang, Jewon Yeom, Juan Yeo, Hyunggyu Lim, and Taesup Kim. Stable on-policy distillation through adaptive target reformulation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 42217–42227, 2026. doi: 10.18653/v1/2026.findings-acl.2094. URL https://aclanthology.org/2026.findings-acl.2094/.

Nan Jia, Haojin Yang, Xing Ma, Jiesong Lian, Shuailiang Zhang, Weipeng Zhang, Ke Zeng, Xunliang Cai, and Zequn Sun. Asymmetric on-policy distillation: Bridging exploitation and imitation at the token level. *arXiv preprint arXiv:2605.06387*, 2026. URL https://arxiv.org/abs/2605.06387.

Woogyeol Jin, Taywon Min, Yongjin Yang, Dennis Wei, Yi Zhou, Swanand Ravindra Kadhe, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. URL https://arxiv.org/abs/2603.07079v3.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. VinePPO: Refining credit assignment in RL training of LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29557–29590, 2025. URL https://proceedings.mlr.press/v267/kazemnejad25a.html.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards streamlined distillation for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24872–24895, 2024. URL https://proceedings.mlr.press/v235/ko24c.html.

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*, 2026. URL https://arxiv.org/abs/2606.15912.

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.

Alexander Lin, Jeremy Wohlwend, Howard Chen, and Tao Lei. Autoregressive knowledge distillation through imitation learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 6121–6133, 2020. doi: 10.18653/v1/2020.emnlp-main.494. URL https://aclanthology.org/2020.emnlp-main.494/.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. URL https://thinkingmachines.ai/blog/on-policy-distillation/.

Chiyu Ma, Shuo Yang, Kexin Huang, Jinda Lu, Haoming Meng, Shangshang Wang, Bolin Ding, Soroush Vosoughi, Guoyin Wang, and Jingren Zhou. FIPO: Eliciting deep reasoning with future-KL influenced policy optimization. *arXiv preprint arXiv:2603.19835*, 2026. URL https://arxiv.org/abs/2603.19835.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.06732.

Stephane Ross and Drew Bagnell. Efficient reductions for imitation learning. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 661–668, 2010. URL https://proceedings.mlr.press/v9/ross10a.html.

Stephane 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*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

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.

Bing Shao, Jiazheng Zhang, Long Ma, Yujiong Shen, Senjie Jin, Xin Guo, Yuming Yang, Mingxu Chai, Zhiheng Xi, Boyang Liu, Junlin Shang, Tao Gui, Qi Zhang, and Xuanjing Huang. A token-level analysis of sampled-token reverse-KL on-policy distillation. *arXiv preprint arXiv:2608.25643*, 2026. URL https://arxiv.org/abs/2608.25643.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.03768.

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.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, 2024. doi: 10.18653/v1/2024.acl-long.510. URL https://aclanthology.org/2024.acl-long.510/.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. ScienceWorld: Is your agent smarter than a 5th grader? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11279–11298, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.775. URL https://aclanthology.org/2022.emnlp-main.775/.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. TIP: Token importance in on-policy distillation. *arXiv preprint arXiv:2604.14084*, 2026. URL https://arxiv.org/abs/2604.14084.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. WebShop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems*, volume 35, pp. 20744–20757, 2022. doi: 10.52202/068431-1508. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/82ad13ec01f9fe44c01cb91814fd7b8c-Abstract-Conference.html.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.03629.

Binbin Zheng, Xing Ma, Yiheng Liang, Jingqing Ruan, Xiaoliang Fu, Kepeng Lin, Benchang Zhu, Ke Zeng, and Xunliang Cai. SCOPE: Signal-calibrated on-policy distillation enhancement with dual-path adaptive weighting. *arXiv preprint arXiv:2604.10688*, 2026. URL https://arxiv.org/abs/2604.10688.

Qiyong Zhong, Mao Zheng, Mingyang Song, Xin Lin, Jie Sun, Yiqi Zhao, Xiaodi Wang, 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.

A Optimization details

The paired comparison in Equation 11 assumes successful replay to the same state and history before the selected response. With independent environment instances, a common continuation policy, and equal remaining horizons, the comparison conditions are

$$\begin{aligned}(s^S_{k^\star}, c^S_{k^\star}) = (s^T_{k^\star}, c^T_{k^\star}) = (s_{k^\star}, c_{k^\star}),\\ p^S_{\mathrm{cont}} = p^T_{\mathrm{cont}} = p_\circ, \qquad B^S_{\mathrm{rem}} = B^T_{\mathrm{rem}}.\end{aligned} \tag{18}$$

As in Section 3.3, T identifies the use of the teacher response, which replaces the student's original response in the assistant history. Both continuations share decoding settings and seed schedules. Replay or runtime errors exclude a trajectory from $\mathcal{D}_{\mathrm{pair}}$. The positive part of the final task outcome difference, $[\Delta_\zeta]_+$, equals the indicator g_ζ for additional teacher supervision in Equation 13.

For a fixed batch $\mathcal{D}$ of the student's own trajectories, vanilla OPD and OG-OPD apply the same clipped policy surrogate to the same valid original response tokens. Only the turn weights differ. The valid token set and normalization are

$$\mathcal{M}_{\mathrm{orig}}(\zeta) = \{(k, j) : j \in \mathcal{M}_k(\zeta)\}, \qquad Z = \sum_{\zeta \in \mathcal{D}} |\mathcal{M}_{\mathrm{orig}}(\zeta)|, \tag{19}$$

where $\mathcal{M}_k(\zeta)$ excludes prompts, environment observations, padding, and positions without valid teacher scores. The following identities concern batches with $Z > 0$. For importance ratio ρ and fixed signed coefficient a representing token-level teacher supervision, the shared surrogate is

$$\ell_{\mathrm{ratio}}(\rho, a) = \max\{-\rho a, -\operatorname{clip}(\rho, 1 - \varepsilon_-, 1 + \varepsilon_+)a\}\,, \tag{20}$$

$$\ell_{\mathrm{pol}}(\rho, a) = \begin{cases}\min\{\ell_{\mathrm{ratio}}(\rho, a), -\kappa a\}, & a < 0,\\ \ell_{\mathrm{ratio}}(\rho, a), & a \geq 0.\end{cases} \tag{21}$$

Here $\varepsilon_-, \varepsilon_+ > 0$ are clipping widths and $\kappa > 1$ bounds the surrogate for $a < 0$. Token-level teacher supervision retains the signed coefficients $\psi_{k,j}$ defined in Section 2.2. Recorded probabilities, turn weights, and final task outcome indicators are held fixed during differentiation. In implementation, the surrogate uses the numerically bounded ratio

$$\widehat{\rho}_{k,j}(\theta) = \exp(\operatorname{clip}(\log p_\theta(z_{k,j} \mid v_{k,j}) - \log p_\circ(z_{k,j} \mid v_{k,j}), -C, C))\,, \tag{22}$$

where $C = 20$. The bounded ratio equals Equation 3 when clipping is inactive. Equation 24 holds for either ratio. The following identities apply to the fixed-coefficient policy surrogate, not the exact reverse-KL objective at arbitrary student parameters.

All valid original tokens at a turn share its weight, applied through the recorded advantage

$$a^{\mathrm{OG-OPD}}_{k,j} = \mathrm{sg}[\omega_k \psi_{k,j}]\,. \tag{23}$$

The clipped loss is positively homogeneous in its signed argument. For fixed $\omega \geq 0$,

$$\ell_{\mathrm{pol}}(\rho, \omega a) = \omega \ell_{\mathrm{pol}}(\rho, a). \tag{24}$$

Multiplication by a positive fixed weight preserves the sign of a and the ordering of the arguments to each minimum and maximum; both sides vanish for $\omega = 0$. On this fixed batch, weighting the recorded advantage is therefore equivalent to weighting the original turn loss $\mathcal{Q}_k$. Substituting Equation 14 yields the decomposition in Equation 15.

Under Equation 14, $\gamma_\zeta = 0$ when $\mathrm{g}_\zeta = 0$ or $\beta_{k^\star} \geq \beta_\uparrow$. With no additional weight, the objective reduces to $\mathcal{L}_{\mathrm{base}}$. Unit pre-calibration weights further recover Equation 5. For fixed recorded coefficients, positive turn weights rescale each token's loss gradient without reversing its direction. Reweighting across turns can nevertheless change the direction of the batch gradient.

Gradient decomposition and weight bounds. For the fixed batch and recorded weights above, differentiation can be taken inside the finite token sums:

$$\nabla_\theta \mathcal{L}_{\mathrm{OG-OPD}} = \frac{1}{Z} \sum_{\zeta \in \mathcal{D}} \sum_{k \in \mathcal{V}(\zeta)} \omega_k \sum_{j \in \mathcal{M}_k(\zeta)} \nabla_\theta \ell_{\mathrm{pol}}(\rho_{k,j}(\theta), \mathrm{sg}[\psi_{k,j}])\,. \tag{25}$$

Splitting ω_k into its base and additional terms recovers Equation 17, with no gradient through the sampled final task outcomes. Wherever the derivatives exist, the triangle inequality gives

$$\|\nabla_\theta \mathcal{L}_{\mathrm{OG-OPD}} - \nabla_\theta \mathcal{L}_{\mathrm{base}}\| \leq \frac{1}{Z} \sum_{\zeta \in \mathcal{D}_{\mathrm{pair}}} \gamma_\zeta \|\nabla_\theta \mathcal{Q}_{k^\star}(\theta; \zeta)\|\,. \tag{26}$$

This expression bounds the additional loss gradient in terms of the selected turn gradients and their weights. A small number of upweighted turns alone does not imply a small gradient change, and the bound does not establish an improvement in task success after optimization.

The floor rule provides explicit bounds on the coefficients. Since $0 < \beta_k \le \beta_{\max}$ on valid turns and $\mathsf{g}_\zeta \in \{0, 1\}$,

$$0 \le \gamma_\zeta \le \beta_\uparrow, \qquad \beta_k \le \omega_k \le \max\{\beta_{\max}, \beta_\uparrow\}. \tag{27}$$

These are bounds on supervision coefficients, not on gradient norms. In particular, no additional normalization by the sum of weights is introduced, so the original valid token denominator Z is preserved.

Interpretation of outcome-based calibration. For a restored state s, history c, response z, and frozen current student, define the probability of a successful final task outcome by

$$V_{p_\circ}(s, c, z) = \mathbb{E}_\varsigma[\mathsf{S}(\mathcal{U}(s, c, z; p_\circ, \varsigma))]\,. \tag{28}$$

For fixed $(s, c, z^S, z^T, p_\circ)$, correctly restored states and the intended sampling marginals give

$$\mathbb{E}_\varsigma\Big[\mathsf{S}(\tilde{\zeta}^T) - \mathsf{S}(\tilde{\zeta}^S)\Big] = V_{p_\circ}(s, c, z^T) - V_{p_\circ}(s, c, z^S). \tag{29}$$

The shared seed schedule couples the final task outcomes without changing this identity. Under these fixed conditions,

$$\begin{aligned} \mathbb{E}_\varsigma[\Delta_\zeta] &= \Pr_\varsigma(\Delta_\zeta = 1) - \Pr_\varsigma(\Delta_\zeta = -1), \\ \mathbb{E}_\varsigma[\mathsf{g}_\zeta] &= \Pr_\varsigma(\Delta_\zeta = 1). \end{aligned} \tag{30}$$

The first expectation measures net execution benefit; the second measures teacher rescue probability under paired sampling. OG-OPD uses one observed teacher rescue to condition additional weight, rather than estimating net expected benefit from repeated trials. These execution statistics do not establish the effect of an individual training update.

For a selected turn with fixed pre-calibration weight, let $h_\zeta = [\beta_\uparrow - \beta_{k^\star}]_+$ and $p_+ = \Pr_\varsigma(\Delta_\zeta = 1)$ under the same fixed comparison conditions. Since $\gamma_\zeta = h_\zeta \mathsf{g}_\zeta$, its first two moments are

$$\mathbb{E}_\varsigma[\gamma_\zeta] = h_\zeta p_+, \qquad \mathrm{Var}_\varsigma(\gamma_\zeta) = h_\zeta^2 p_+(1 - p_+). \tag{31}$$

The expected additional weight therefore depends on both the probability of an observed rescue and the distance to the weight floor. When $h_\zeta = 0$, even a positive paired outcome leaves the weight unchanged. For $h_\zeta > 0$, the probability of an actual increase equals p_+ under these fixed conditions.

B Empirical analysis details

Section 3.1 introduces the empirical analysis with a fixed Qwen3-1.7B student and Qwen3-32B teacher on ScienceWorld and WebShop. Continuations S and TS use the current student after the student's original action and the teacher action, respectively; TT uses the teacher after the same teacher action to compare what the teacher and current student can complete. Eligible positions within the student's own trajectories are sampled uniformly without replacement, independently of OG-OPD's candidate selection rule. Under a bounded proposal budget, we retain the first replay-valid sampled position with a teacher response whose parsed action is legal and differs from the student's original action. The empirical analysis therefore includes only decisions with valid recorded scores, successful replay, and valid teacher responses.

Each retained decision is evaluated in five matched trials. Let $\tilde{\zeta}_m^b$ be continuation b in trial m. With M valid trials, we compute each continuation's success frequency and how often the paired student continuations differ in final task outcome:

$$f_b = \frac{1}{M}\sum_{m=1}^{M} \mathsf{S}(\tilde{\zeta}_m^b), \quad b \in \{S, TS, TT\}, \qquad q_{\text{diff}} = \frac{1}{M}\sum_{m=1}^{M} |\mathsf{S}(\tilde{\zeta}_m^{TS}) - \mathsf{S}(\tilde{\zeta}_m^S)|. \tag{32}$$

We assign outcome categories using lower and upper success frequency thresholds $\alpha_- < \alpha_+$:

$$\begin{aligned} \text{benign gaps}: &\quad f_S \ge \alpha_+,\ f_{TS} \ge \alpha_+, \\ \text{teacher rescue}: &\quad f_S \le \alpha_-,\ f_{TS} \ge \alpha_+, \\ \text{teacher harm}: &\quad f_S \ge \alpha_+,\ f_{TS} \le \alpha_-, \\ \text{teacher-only completion}: &\quad f_{TT} \ge \alpha_+,\ f_{TS} \le \alpha_-, \\ \text{outcome-critical}: &\quad q_{\text{diff}} \ge \alpha_+. \end{aligned} \tag{33}$$

Here $M = 5$, $\alpha_- = 1/5$, and $\alpha_+ = 3/5$, corresponding to at most one and at least three successes. Benign gaps use each action's success frequency, whereas outcome-critical decisions require different final task outcomes in at least three matched pairs. Categories can overlap, including teacher harm and teacher-only completion. Model parameters remain fixed, and all three continuations use matched seed schedules, decoding settings, and remaining interaction budgets. Teacher responses replace both the executable action and the assistant response retained in history.

The retained set contains 329 of 330 selected ScienceWorld decisions after excluding one replay failure, and all 274 selected WebShop decisions. We divide gap magnitudes into five groups using empirical quantiles of χ_k in Equation 6; large and small gaps refer to the highest and lowest groups. Boundaries are computed separately over all eligible positions in each benchmark, including unselected positions. In ScienceWorld, the large- and small-gap groups contain 172 and 37 decisions, respectively; the corresponding counts for WebShop are 73 and 33. Empirical analysis quantiles give equal weight to each eligible position, whereas the training quantiles in Equation 36 balance trajectories. Decisions with completed trials are assigned to these existing groups, yielding unequal group sizes.

Counts and cross-benchmark aggregation. Table 4 gives the counts behind Figure 2. For outcome category j, let $I_j(d)$ equal one when decision d meets its criterion and zero otherwise. Let $\mathcal{D}_{j,e}$ contain the decisions in the corresponding gap group of benchmark e. We divide the number of decisions in the category by the group size, then average the two benchmark percentages with equal weight:

$$r_{j,e} = \frac{100}{|\mathcal{D}_{j,e}|} \sum_{d \in \mathcal{D}_{j,e}} I_j(d), \qquad \bar{r}_j = \frac{r_{j,\text{ScienceWorld}} + r_{j,\text{WebShop}}}{2}. \tag{34}$$

The calculation applies to all five categories, with percentages averaged before rounding. Repeated trials determine category membership; each decision contributes once to every category it meets. Rates use the number of retained decisions in the corresponding gap group as their denominator.

Confidence intervals. Within each benchmark and gap group, we resample retained decisions with replacement 2,000 times and take the 2.5th and 97.5th percentiles to obtain the 95% confidence intervals in Figure 2. Each decision is classified using its five matched trials before resampling. The zero observed count for small-gap outcome-critical decisions on ScienceWorld produces a degenerate bootstrap interval, not evidence of a zero population proportion.

Table 4: **Counts underlying the empirical analysis.** Each benchmark entry gives the number of decisions in an outcome category divided by the corresponding gap group size, followed by the percentage in parentheses. Categories follow Equation 33. The average gives equal weight to the two benchmarks.

Outcome category	Gap group	ScienceWorld	WebShop	Average
Benign gaps	Large	2/172 (1.2%)	7/73 (9.6%)	5.4%
Teacher rescue	Large	9/172 (5.2%)	9/73 (12.3%)	8.8%
Teacher harm	Large	4/172 (2.3%)	14/73 (19.2%)	10.8%
Teacher-only completion	Large	26/172 (15.1%)	19/73 (26.0%)	20.6%
Outcome-critical	Small	0/37 (0.0%)	5/33 (15.2%)	7.6%

B.1 Ranking by Gap Magnitude

We further examine whether larger gaps preferentially identify decisions where teacher guidance benefits the current student. For Figure 5, let $\mathcal{S}_k$ contain the top k of n retained decisions ranked by decreasing gap magnitude. With teacher rescue and teacher harm indicators R_j and H_j from Equation 33, the plotted measures are

$$\text{coverage}(k) = \frac{k}{n}, \qquad \text{precision}(k) = \frac{\sum_{j \in \mathcal{S}_k} R_j}{k}, \qquad \text{net}(k) = \frac{\sum_{j \in \mathcal{S}_k} (R_j - H_j)}{k}. \tag{35}$$

Ties are broken deterministically by task identity. WebShop's net rescue is negative over much of the coverage range, while ScienceWorld's stays near zero or slightly positive. These curves compare the frequencies of the decision categories in Equation 33, not the expected outcome difference in Equation 29. They describe the retained decisions under the frozen current student, with small selected sets at low coverage.

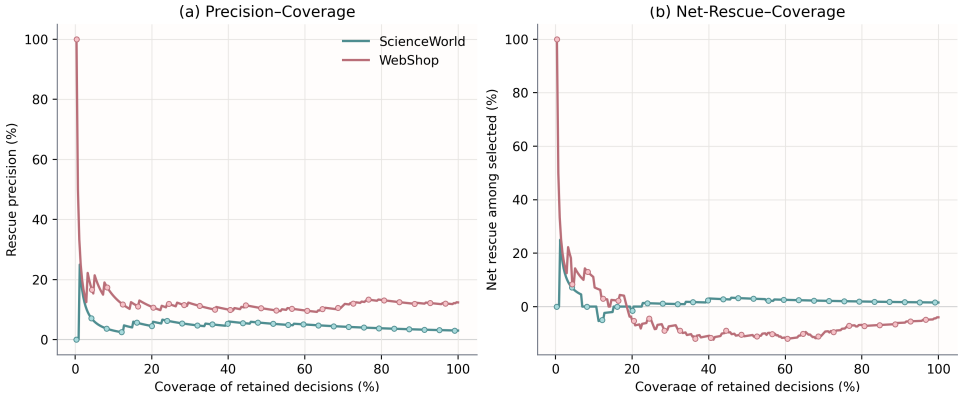


Figure 5: **Final task outcomes under gap magnitude ranking.** Decisions are ranked by decreasing gap magnitude. Rescue precision (left) is the teacher rescue rate; net rescue (right) subtracts the teacher harm rate.

C EXECUTION CONTRACT

ALFWorld, ScienceWorld, and WebShop responses are parsed into their respective household commands, scientific actions, and search or click operations. The environment receives the parsed action, while the complete assistant response remains in the interaction history. For the table comparisons, methods within each benchmark share evaluation prompts, response token budgets, history retention, action parsing, interaction horizons, termination handling, and score extraction. Action parsing follows each benchmark's shared protocol, without additional repair rules for individual methods.

Teacher and student log probabilities are evaluated on the same valid original response tokens, including generated reasoning and action text but excluding prompts, observations, and padding. Gap magnitudes from Equation 6 determine trajectory-relative weighting and candidate selection, while the signed coefficients in Equation 3 determine token-level teacher supervision. Missing or nonfinite scores are excluded. Turns with empty valid token masks have no gap or weight and are excluded from the loss and threshold populations; trajectories with no valid turns contribute neither supervision nor candidates. A first valid turn $k_0 > 0$ supplies the weighting reference but has no valid change from the preceding turn and cannot use the candidate exception reserved for $k = 0$.

For the candidate thresholds, let $\mathcal{H}_\zeta^f$ contain eligible positions for statistic $f \in \{\nu, \upsilon\}$ and let $\mathcal{D}_f$ contain trajectories with nonempty $\mathcal{H}_\zeta^f$. Only positive valid changes $\upsilon_k > 0$ enter the population for υ. The empirical distribution and quantile are

$$\widehat{F}_f(v) = \frac{1}{|\mathcal{D}_f|} \sum_{\zeta \in \mathcal{D}_f} \frac{1}{|\mathcal{H}_\zeta^f|} \sum_{k \in \mathcal{H}_\zeta^f} \mathbb{I}[f_k \le v], \qquad \vartheta_f = \inf\{v : \widehat{F}_f(v) \ge q_f\}. \tag{36}$$

Here $q_f \in (0, 1)$ is the quantile level. Each contributing trajectory receives equal total weight in the threshold population. Since $\log(\varepsilon_0 + \chi)$ is strictly increasing, the implementation equivalently transforms the inverse CDF quantile of the gap magnitude. If the threshold population is insufficient, pre-calibration weights are retained.

Teacher responses use the native prompt and parser for each benchmark with native thinking mode disabled. Each complete response, including any visible rationale, replaces the student's original response in continuation history. Empty or unparseable responses and those yielding inadmissible or unchanged actions are excluded. Within a fixed proposal budget, candidates are checked chronologically. The first replay-valid turn with a legal teacher action different from the student's original action is selected for paired student continuations. Independent environment instances replay the task prefix and check the intended observation and action context. Both continuations use the same frozen current student, remaining budget, decoding settings, and seed schedule at each step. Replay and runtime failures invalidate the comparison rather than count as task failures.

Only tokens from the student's own trajectories enter the loss. The frozen rollout version supplies the importance ratio denominator, and asynchronous collection retains the existing staleness control. Both paired student continuations use that same student version. Their final task outcomes determine

Table 5: **OG-OPD weighting and calibration settings.** Candidate thresholds follow Equation 36. Training uses one matched pair per selected position, whereas the empirical analysis uses five matched trials per decision.

Parameter	Value
Pre-calibration weight cap $\beta_{\max}$	1.2
Positive-outcome weight floor $\beta_{\uparrow}$	1.5
Gap-magnitude / positive-increase quantile	$q_{\nu} = 0.5$ / $q_{v} = 0.6$
Candidate checks / teacher proposals per trajectory	At most 2 / 2
Paired-evaluation positions per trajectory	At most 1
Matched trials per selected position	1 pair

the additional weight in Equation 14; the original token-level teacher supervision remains the signed coefficient in the loss.

D Experimental details

D.1 Model configurations and comparison scope

The model pairings follow Section 4.1. Runs use one node with $8\times$A800 GPUs. The empirical analysis, gap magnitude ranking, ablations, and timing use the Qwen3-32B teacher and Qwen3-1.7B student. The WebShop training dynamics in Section 4.2 cover both configurations.

Table 5 summarizes the OG-OPD weighting and calibration settings. Complete training configurations are provided with the supplementary source code.

Figure 4 reports WebShop training curves separately from table results averaged over three seeds. Entropy and Teacher NLL are averaged over five steps, and the remaining rates are pooled over five batches.

The RL teachers are obtained by training Qwen3-8B separately on ALFWorld and WebShop with GiGPO (Feng et al., 2025), following the teacher training approach of Chen et al. (2026). Both use AdamW with learning rate 10^{-6} and weight decay 0.01. The discount factor is 0.95, the rollout group size is 8, and the invalid action penalty coefficient is 0.1. Training batch sizes are 32 for ALFWorld and 16 for WebShop, with a PPO minibatch size of 64. Teacher training uses a prompt length limit of 10,240 tokens, with interaction limits of 30 steps for ALFWorld and 15 steps for WebShop. Responses are limited to 512 tokens, with thinking mode disabled. Training and validation temperatures are 1.0 and 0.4, respectively. We use the ALFWorld checkpoint at step 150 and the WebShop checkpoint at step 100, held fixed across all compared methods.

Performance and ablation comparisons evaluate complete training procedures at step 200. Ablations retain each procedure's resulting number of additionally upweighted turns and total supervision weight, rather than matching these quantities across variants. OG-OPD includes additional environment execution and final task outcome feedback, so these comparisons assess the complete algorithm. They use a fixed number of optimization steps, with actual training costs reported separately in Appendix E.

D.2 Evaluation tasks, metrics, and aggregation

Following Wang et al. (2026), ALFWorld evaluation uses 140 Seen tasks in environments encountered during training and 134 Unseen tasks with room layouts and object combinations not encountered during training. We also use TCOD's fixed Hard collection of 121 tasks drawn from the training split, on which its teacher failed in all ten sampling attempts. We report SR for each collection and Overall SR over all 395 tasks pooled together. ScienceWorld evaluates 1,308 canonical tasks. WebShop uses 100 held-out sessions, as in Chen et al. (2026). Interaction limits are 30 steps for ALFWorld and ScienceWorld and 15 steps for WebShop. ScienceWorld success requires full completion. WebShop uses the shared TCOD success criterion of terminal reward above 0.5.

For episode j, let $s_{j,t}$ be its ScienceWorld environment score on a 0–100 scale and $r_{j,\mathrm{end}}$ its WebShop terminal reward on a 0–1 scale. Score is computed for each benchmark as

$$\mathrm{Score}_{\mathrm{SW}} = \frac{1}{n}\sum_{j=1}^{n} \mathrm{clip}\Big(\max_{t} s_{j,t}, 0, 100\Big), \quad \mathrm{Score}_{\mathrm{WS}} = \frac{100}{n}\sum_{j=1}^{n} r_{j,\mathrm{end}}. \tag{37}$$

Rounds averages the actual interaction lengths of all evaluated episodes, including failures. Score is extracted within each episode as defined above, using the maximum ScienceWorld score and the terminal WebShop reward.

For episode success label Y_j under the criterion for that benchmark and executed interaction length L_j, the remaining aggregates are

$$\mathrm{SR} = \frac{100}{n}\sum_{j=1}^{n} Y_j, \qquad \mathrm{Rounds} = \frac{1}{n}\sum_{j=1}^{n} L_j, \qquad \mathrm{Overall}_{\mathrm{ALF}} = \frac{\sum_{s\in\mathcal{S}} n_s\,\mathrm{SR}_s}{\sum_{s\in\mathcal{S}} n_s}, \tag{38}$$

where $\mathcal{S} = \{\mathrm{Seen}, \mathrm{Unseen}, \mathrm{Hard}\}$ and n_s is the number of evaluated tasks in collection s. Overall SR gives each task equal weight.

Trained-model results in the main and ablation tables use the fixed endpoint at optimizer step 200 and report means and sample standard deviations across three seeds. For any evaluation metric M, we first compute the metric over the benchmark's episodes for each seed, then report

$$\overline{M}_{200} = \frac{1}{3}\sum_{r=1}^{3} M_{r,200}, \qquad s_{M,200} = \sqrt{\frac{1}{3-1}\sum_{r=1}^{3}\left(M_{r,200} - \overline{M}_{200}\right)^2}. \tag{39}$$

Each trained model is evaluated at the fixed endpoint, and zero-shot references are evaluated without training. Timing is measured per run and averaged across seeds as described in Appendix E.

E TRAINING EFFICIENCY

Table 6: **Training cost of vanilla OPD and OG-OPD.** Runs use 8×A800 GPUs with the Qwen3-32B teacher and Qwen3-1.7B student. OG-OPD timing includes outcome-based calibration, and Cost is normalized to vanilla OPD. Rounds reports mean evaluation interaction rounds. Boldface marks the best value per benchmark.

Benchmark	**Method**	Rounds ↓	s/step ↓	Cost ↓
ScienceWorld	Vanilla OPD	12.6	**156.97**	1.00×
	OG-OPD	**11.6**	223.58	1.42×
WebShop	Vanilla OPD	6.8	**67.21**	1.00×
	OG-OPD	**5.6**	99.70	1.48×

For each run, elapsed time from training start through step 200 is divided by 200, then averaged across three seeds. The measurement includes state replay, teacher response generation, and paired student continuations, and excludes final synchronization and evaluation. Deployment uses only the trained student.

Let $T_{m,e,r}$ denote this measured elapsed time for method m, benchmark e, and training seed r. With $K = 200$ optimization steps and $R = 3$ seeds, the reported seconds per step and normalized cost are

$$\bar{\tau}_{m,e} = \frac{1}{R}\sum_{r=1}^{R}\frac{T_{m,e,r}}{K}, \qquad \mathrm{Cost}_{m,e} = \frac{\bar{\tau}_{m,e}}{\bar{\tau}_{\mathrm{OPD},e}}. \tag{40}$$

Cost is the ratio of the two mean times per step, rather than a mean of ratios computed for individual seeds. The corresponding absolute and relative increases are

$$\Delta\bar{\tau}_e = \bar{\tau}_{\mathrm{OG-OPD},e} - \bar{\tau}_{\mathrm{OPD},e}, \qquad \eta_e = 100\left(\mathrm{Cost}_{\mathrm{OG-OPD},e} - 1\right). \tag{41}$$

Using the rounded times in Table 6, these increases are 66.61 seconds per step on ScienceWorld and 32.49 seconds on WebShop, or approximately 42% and 48%. They describe end-to-end training overhead under the reported hardware and execution settings, not a separate timing measurement of each calibration operation.