CAPABLE TEACHER, MYOPIC SCORER: REWARD-FREE TEACHER REPAIR FOR AGENTIC ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Agentic on-policy distillation (OPD) trains a student agent on its own trajectories while a teacher scores every token. A natural explanation of where it goes wrong is compounding error: the student's mistakes carry its trajectories beyond the region where the teacher is reliable, so recent methods keep training inside that region, through episode-length curricula, teacher-guided turns, prefix replay, or teacher intervention where the two disagree. We show this explanation is incomplete. From the states where the student departs most from the teacher, the teacher's own takeovers still finish the task, yet the student trained on its token scores keeps running out of turns. We attribute this to the myopia of next-token scoring. The teacher rates each token without seeing where the episode ends, so its scores let through the safe, delaying moves that outcome-based RL would punish with zero reward, a penalty agentic OPD lacks. We call this difference between what the teacher does and what it scores the *rollout–scorer gap*. We propose Rollout-to-Scorer Distillation (R2S), which repairs the scorer before distillation by mining high-disagreement states from the untrained student, running teacher takeovers from them, keeping the states where enough takeovers pass a reward-free check such as consensus or an LLM judge, and fine-tuning the teacher to give its takeover-informed scores on every assistant token from the divergence turn onward, without the takeover in context. The repaired teacher is frozen as the sole scorer for standard OPD, requiring no task reward during repair or training. Across three environments, two teachers and four students, R2S outperforms three state-of-the-art agentic-OPD methods in all twelve teacher–student–environment cells, by 12.8 success points on average over the best of them in each cell. We release our code in the supplementary material.

1 INTRODUCTION

Agentic on-policy distillation (OPD) trains an agent on its own trajectories. The student acts in a live environment, and a stronger teacher scores every token it produces (Liao et al., 2026). The teacher never acts in this loop; it only scores, and its next-token distribution on the student's histories is the whole training signal. This form of dense supervision also underlies the multi-teacher post-training pipelines of MiMo-V2-Flash and Nemotron-Cascade 2 (Xiao et al., 2026; Yang et al., 2026).

A natural explanation of where agentic OPD goes wrong is compounding error. The student's mistakes carry its trajectories beyond the region where the teacher is reliable, and its scores there are noise. Recent agentic-OPD methods act on this reading and keep training inside the trusted region, by shortening the student's episodes (Wang et al., 2026), handing turns to the teacher (Li et al., 2026a), replaying teacher prefixes (Liao et al., 2026), or intervening where the two disagree (Chen et al., 2026), and report gains in their own settings.

We show this explanation is incomplete. We take the states where the untrained student disagrees most with the teacher, late in its episodes, the states that compounding error should place furthest outside the teacher's region, and let the teacher act from them. Its takeovers still finish the task. In every environment they beat the student's own outcome from the same states, and they rescue many of the states the student went on to lose. The student has not drifted beyond what the teacher can

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.

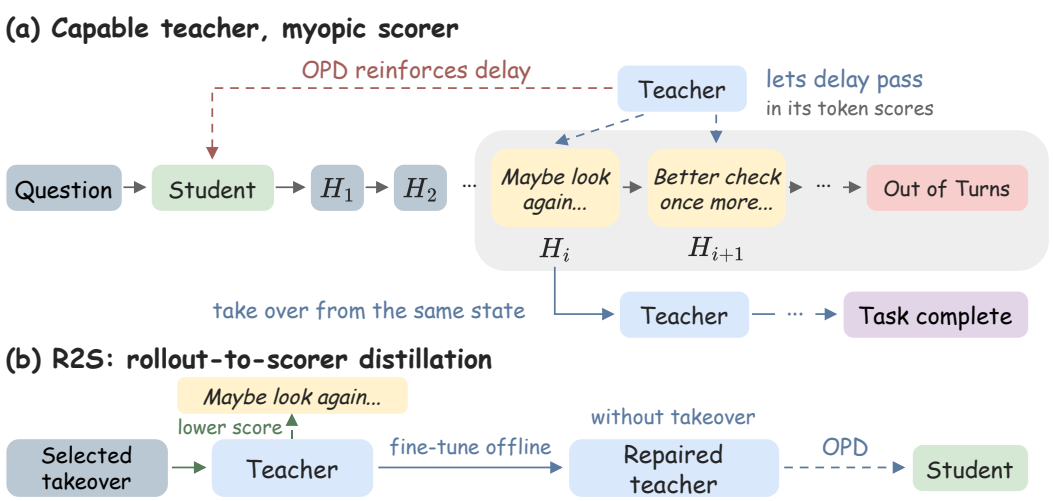

Figure 1: **OPD can reinforce the delays its teacher lets pass.** (a) The teacher's scores let delaying thoughts pass and the episode runs out of turns; its own takeover from that state finishes. (b) With the takeover in context, the teacher scores the delay lower; R2S trains a copy to do so without it.

handle. Yet the teacher's plain scores let the student's delay pass: only with one of its finishing takeovers in context do they mark down the student's hedging and searching, and the student trained on the plain scores keeps running out of turns. Figure 1 illustrates this.

We attribute this to the myopia of next-token scoring. Acting, the teacher writes the whole episode itself: each action is followed by its consequences and by the teacher's own next action, and the episode runs to an outcome. Scoring, it rates each of the student's tokens one at a time, on a history it did not write, without seeing where the episode ends. Under that uncertainty another look or another search keeps options open and is reasonable on its own, so the teacher's scores let it pass. Outcome-based RL punishes such delay with zero reward when the episode runs out, while agentic OPD has no such penalty. We call this difference between what the teacher does and what it scores the *rollout–scorer gap*. This motivates informing the scores with selected teacher continuations.

Rollout-to-Scorer Distillation (R2S) repairs the scorer before distillation. First, the teacher takes over from the untrained student's highest-disagreement states. Second, we keep the states where enough takeovers pass a reward-free check, consensus among them or an LLM judge, and pick the shortest passing one as the reference. Third, we fine-tune the teacher to give its takeover-informed scores on every assistant token from the selected divergence turn onward, without the takeover in context. The repaired teacher is frozen for standard OPD. Student training retains the ordinary OPD objective and rollout recipe, and no task reward enters repair or training. We evaluate R2S across three environments, two teachers and four students under a fixed final-checkpoint protocol. R2S outperforms three state-of-the-art agentic-OPD methods in every cell. The gain over the best of them in each cell averages $+17.0$ points from the 9B teacher and $+8.5$ from the 30B-A3B teacher.

Our contributions are: (1) **The rollout–scorer gap.** Section 3 shows that the teacher's takeovers still finish from the states where the student departs most from it, while its plain scores let the student's delay pass; only a finishing takeover in context turns them against it. (2) **R2S.** Section 4 introduces a reward-free repair that moves the teacher's rollout competence into the scores OPD consumes, then freezes the teacher for standard OPD. (3) **Evaluation.** We evaluate R2S against three state-of-the-art agentic-OPD methods across twelve model–environment cells under a shared protocol, and measure target fidelity, observed failures and completion separately in a long-horizon analysis.

2 Related Work

On-policy distillation for agents. ReOPD identifies a prefix trap: histories closer to the student's distribution can make the teacher's targets less reliable, motivating a curriculum over replayed teacher prefixes (Liao et al., 2026). Our compared methods address related trajectory-level difficulties: TCOD-F2B schedules rollout depth, Guided-OPD allocates teacher-controlled turns, and

FutureBridge validates teacher bridges from high-disagreement states (Wang et al., 2026; Li et al., 2026a; Chen et al., 2026). These methods retain the original scorer. R2S repairs its token scores before student training, a change that can be combined with student-side curricula.

Context and self-distillation. SDFT and Self-Distilled Reasoner use demonstration-conditioned self-teachers to supervise student rollouts (Shenfeld et al., 2026; Zhao et al., 2026). SDPO instead conditions on textual feedback or successful attempts (Hübotter et al., 2026), while OPCD internalizes experiential knowledge and system prompts through on-policy context distillation (Ye et al., 2026). CANON selects privileged solutions by consensus without gold labels (Gkountouras et al., 2026). These works establish that useful distillation signals need not come from task rewards. R2S uses teacher takeovers to repair an external scorer, then freezes it for a separate student's OPD.

Teacher quality and repair. Li et al. (2026c) show that a higher-scoring teacher need not transfer better: compatible reasoning patterns and new capabilities matter. Kaur et al. (2026) find that privileged context can suppress self-correction and degrade thinking models. These results motivate checking what the teacher's local signal teaches, beyond its standalone performance. Distillation Traps and Guards directly calibrates an external teacher with utility and calibration rewards (Zhan et al., 2026). R2S instead fits takeover-informed token targets without task reward and evaluates the repaired scorer through target fidelity and downstream completion. Appendices A.1 and A.2 give the detailed comparisons and extended discussion.

3 Analysis

3.1 Preliminaries: Agentic On-Policy Distillation

Agentic on-policy distillation (OPD) trains a student π_θ to match a teacher T. The student acts in the environment: at each turn t it reads the history h_t (the task, the observations so far and its own earlier responses) and emits a response y_t; the environment returns the next observation. In vanilla OPD the teacher does not act; it scores every token the student produces, given the same history.

The training signal is the teacher's next-token distribution on the student's own histories. The student minimizes a token-level divergence to it; we use the reverse KL,

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{h\sim\pi_\theta}\, \mathrm{KL}\big(\pi_\theta(\cdot \mid h) \,\|\, T(\cdot \mid h)\big), \tag{1}$$

estimated at the sampled token (Xiao et al., 2026; Li et al., 2026c); our top-k control instead uses the teacher's k most probable next tokens. Other OPD variants use forward KL or Jensen–Shannon divergence (Shenfeld et al., 2026; Hübotter et al., 2026); what follows concerns the per-token scoring, not the choice of divergence. The loss reports how closely the student matches the teacher's scores. Nothing in this objective looks past the current token: the teacher scores h without seeing how the episode ends. The analysis below studies one teacher, Qwen3.5-9B, and its Qwen3.5-4B student.

3.2 Where the Student Drifts, the Teacher Still Finishes

Where the student trains is not the whole story. Recent methods trace what goes wrong in agentic OPD to long horizons. The student's early mistakes compound (Ross et al., 2011) and carry it into states beyond the teacher's reach, where the teacher's scores can no longer guide it, so the student learns the wrong thing there. TCOD-F2B, Guided-OPD and FutureBridge therefore keep training within that reach, by shortening the student's episodes, handing turns to the teacher, or intervening where the two disagree.

The explanation assumes that the states the student drifts into lie beyond what the teacher can handle. If so, the teacher acting from them should fail as well. We test this on the states the explanation points to: the turns at which the untrained student's next-token distribution departs most from the teacher's. From each state the teacher takes over eight times. As the success columns of Table 1 show, in every environment its takeovers finish more often than the student did from the same states. The student has not drifted beyond what the teacher can handle.

Its plain scores let the student's delay pass. Once one of its finishing takeovers is in context, the teacher's scores turn against the student's delay: in the drop columns of Table 1, its probability of the student's hedging and searching words falls two to three times as much as that of the rest of

Table 1: **The teacher finishes from the student's states, but its plain scores let the student delay.** States: highest-disagreement turns from turn five on; *Rescued*: lost states a takeover finishes. *Drop*: in the teacher's probability of the student's words with a finishing takeover in context.

	Success (%)			Drop (%)		
Environment	Student	Takeover	Rescued	Hedging	Searching	Other
ALFWorld	39.9	44.2	45.2	8.2	6.2	3.5
WebShop	6.1	10.2	14.4	8.3	7.7	4.2
ScienceWorld	18.6	27.5	51.7	13.0	16.6	5.0

the student's thinking. The plain scores OPD trains on lack this direction: these states are within the teacher's reach, yet the student learns from the teacher's next-token scores there, not from its rollouts. As a result, the student trained on these plain scores to its final checkpoint still runs out of turns in a third of its held-out ALFWorld and ScienceWorld episodes and in more than half of its WebShop ones, about twice as often as when it trains on the repaired teacher.

Takeaway: recoverable states, cautious student

At the states where the student departs most from the teacher, even late in its episodes, the teacher's own takeovers still finish the task. The teacher's plain scores let the student's delay pass, and the student trained on them keeps running out of turns.

3.3 The Scorer Never Sees Where the Episode Ends

Outcome-based RL punishes delay with zero reward when the episode runs out of turns; plain scoring lacks this continuation signal. Acting, the teacher generates a continuation from the student's state, observes the consequences of its actions, and reaches an outcome. Scoring, it rates the student's tokens given the observed history, without that continuation in context. Its probabilities may encode expectations about future outcomes, but ordinary OPD supplies no explicit feedback from the realized task outcome. This distinction matters when useful information gathering and costly delay look similar locally. Another look or search keeps options open, but it spends one turn of a fixed budget, which the student trained on plain scores often uses up. Plain next-token scores carry no penalty for delay: each extra look is scored on its own, and on its own it is reasonable. Conditioning the scorer on a selected teacher continuation changes its supervision on the same student text. We call the distinction between the teacher's demonstrated rollout competence and its token-level supervision the *rollout–scorer gap*.

Design implications. We mine the untrained student's high-disagreement states and use selected teacher takeovers to inform the scores supplied there. Agreement on an observable outcome or an offline filter supplies the selection rule without reading task reward. We train the teacher to reproduce takeover-informed distributions on every assistant token from the divergence turn onward, without the takeover in context. The repaired teacher then becomes the sole scorer for ordinary OPD.

4 Rollout-to-Scorer Distillation

R2S targets the rollout–scorer gap of Section 3. The teacher's rollouts from the student's states carry the judgment its scores lack. As illustrated in Figure 2, we therefore let the teacher take over from those states, read off the scores it gives with its own ending in view, and fine-tune it to give those scores closed-book, with no takeover in context; the refrozen teacher then scores ordinary OPD. The student never sees a takeover.

The teacher takes over where the student departs most from it. We first let the untrained student play the training tasks. In each of its episodes we mine the divergence turn: the turn where the student's thinking departs most from the teacher's, measured by OPD's own per-token divergence averaged over that thinking,

$$\Delta_t = \frac{1}{|y_t^{\text{think}}|} \sum_{i \in y_t^{\text{think}}} \left[\log \pi_\theta(y_{t,i} \mid h_t, y_{t,<i}) - \log T_0(y_{t,i} \mid h_t, y_{t,<i}) \right], \tag{2}$$

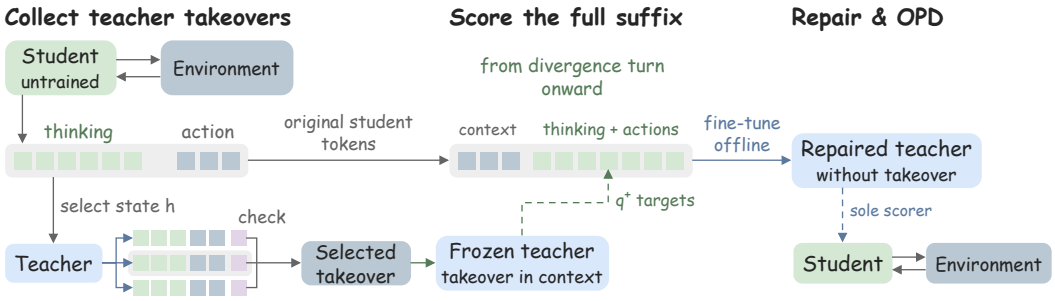

Figure 2: **R2S repairs the teacher's scores on the full assistant suffix before ordinary OPD.** With a selected takeover from the untrained student's state in context, the frozen teacher T_0 scores the student's thinking and actions from the divergence turn onward (q^+). A copy learns these targets without the takeover and becomes the sole OPD scorer.

where y_t^{think} is the thinking of turn t, the part of y_t before the action, and T_0 is the original teacher; the divergence turn has the largest Δ_t. Only the thinking is compared, because it decides the action that follows. From the state at each mined turn, the teacher takes over K times and plays the episode to its end. The choice among these takeovers is general: any check that tells a finishing takeover from a failing one will do, such as consensus among the takeovers, an LLM-as-a-judge, or the ground truth, which is the task's own 0–1 reward. Consensus and a judge need no reward, so R2S can also run where no reward is available. A state is kept when at least m of its takeovers pass the check, and the shortest of them is selected as τ_h.

The judgment supervises the complete suffix from the divergence state. For a kept state h, the frozen teacher T_0 scores the rest of the student's own episode, every token of its thinking and actions from h to the end, with the takeover τ_h in context. These informed scores are the targets q^+. The takeover only informs them; its tokens never replace the student's. A trainable copy T_ϕ sees the same episode without the takeover and learns to give the same scores:

$$\mathcal{L}_{\text{R2S}}(\phi) = \mathbb{E}_h \, \mathbb{E}_{h'} \Big[\text{CE}\big(\underbrace{T_0(\cdot \mid \tau_h, h')}_{q^+}, \; T_\phi(\cdot \mid h') \big) \Big], \tag{3}$$

where h' runs over the student's histories from h to the end of its episode, one per token the student wrote, and CE is the cross-entropy. Refrozen, T_ϕ is an ordinary next-token scorer, and OPD with either estimator uses it in place of T_0.

5 Experiments

Our experiments ask four questions. RQ1: does repairing the scorer improve OPD across teachers, students and environments? RQ2: can the takeovers be selected without reward? RQ3: which parts of R2S are necessary? RQ4: what does the repaired scorer change in long-horizon behavior? We answer these questions below.

5.1 Setup

Teachers and students. Qwen3.5-9B (Qwen Team, 2026) teaches Qwen3.5-2B and Qwen3.5-4B on three environments, and Qwen3.5-4B teaches itself in Table 3. Qwen3-30B-A3B-Instruct-2507 (Yang et al., 2025) teaches Qwen3-4B and Qwen3-8B on the same three environments. We chose these pairs to vary three things: model size, from 2B to 30B parameters; model generation, Qwen3.5 and Qwen3; and architecture, a dense 9B teacher and a mixture-of-experts 30B-A3B teacher. Every prompt carries one fixed worked demonstration per environment; native reasoning mode is off, and each turn opens with one to three sentences of plain-text deliberation before the tool call, as detailed in Appendix D.

Environments. ALFWorld (household tasks), WebShop (shopping) and ScienceWorld (science experiments) are scored by success (%). ALFWorld and ScienceWorld have a seen and an unseen split, and the main text reports their mean; WebShop is evaluated on its 500 canonical goals.

Table 2: **Main results: R2S versus vanilla OPD, TCOD-F2B, Guided-OPD and FutureBridge.** Success (%) of one training run per cell, mean±std over three decodes of its final checkpoint; every split in Table 14. The best and second best of the five trained students in each column are highlighted.

Method	Qwen3.5-9B → Qwen3.5-2B / 4B					
	ALFWorld		WebShop		ScienceWorld	
	2B	4B	2B	4B	2B	4B
Teacher alone	63.1±1.7		33.5±0.3		32.0±1.8	
Student alone	15.1±1.8	43.4±1.0	17.6±1.1	33.7±0.9	3.1±0.4	28.0±1.5
Vanilla OPD	58.0±2.4	65.8±2.6	35.7±1.6	34.6±2.9	23.7±2.0	34.6±3.5
TCOD-F2B	62.7±2.0	59.6±1.5	33.1±0.5	35.1±0.9	28.0±2.6	30.6±0.7
Guided-OPD	63.6±2.2	61.6±2.0	35.6±0.6	35.5±1.3	25.9±3.7	32.2±1.1
FutureBridge	56.8±2.8	58.4±3.7	36.3±1.0	33.9±1.7	19.0±2.0	30.7±1.3
R2S (ours)	77.9±1.3	81.9±2.4	57.9±1.6	55.9±1.4	37.9±0.6	47.8±3.0
Method	Qwen3-30B-A3B-Instruct-2507 → Qwen3-4B / 8B					
	ALFWorld		WebShop		ScienceWorld	
	4B	8B	4B	8B	4B	8B
Teacher alone	42.3±1.1		2.9±0.3		32.9±3.2	
Student alone	30.4±2.3	46.9±0.8	15.5±1.2	23.6±1.2	9.0±1.4	24.0±3.9
Vanilla OPD	41.2±1.7	44.5±2.6	24.7±1.1	20.3±0.9	38.0±1.2	36.8±1.0
TCOD-F2B	42.6±1.1	48.8±1.8	22.8±0.7	23.4±0.7	38.4±0.9	34.9±0.5
Guided-OPD	42.0±0.4	47.0±1.9	23.3±1.6	22.3±0.8	34.7±1.8	33.0±1.5
FutureBridge	41.8±2.6	45.9±1.1	16.7±0.9	20.7±0.7	36.1±1.5	37.1±2.0
R2S (ours)	54.5±1.4	54.7±1.0	28.5±0.6	30.4±1.2	45.8±1.4	50.5±1.4

Protocol. Each teacher–student–environment cell is one 150-step training run. Its final checkpoint is evaluated in three sampled decodes, with no checkpoint selection. The teacher and the untrained student acting alone follow the same prompt and evaluation protocol. In every cell, R2S mines up to 2,400 states, takes $K=8$ takeovers from each and fine-tunes the teacher for one epoch; every other setting is in Appendix C.

Baselines. Vanilla OPD minimizes Eq. (1) with the sampled estimator against the original teacher T_0; R2S runs the same OPD against the repaired teacher T_ϕ, so vanilla OPD is its direct control. The other three are recent agentic-OPD methods, chosen because, like R2S, they train the student by distillation alone, without a task-reward RL objective such as GRPO (Shao et al., 2024). TCOD-F2B (Wang et al., 2026) trains the student on shortened episodes that lengthen over training, Guided-OPD (Li et al., 2026a) hands part of the turns to the teacher, and FutureBridge (Chen et al., 2026), the closest prior method, trains the student on validated teacher bridges from the states where the two disagree most. Their code is public, so all three run under the same models, prompts, estimator and protocol as vanilla OPD and R2S. We leave out methods that combine distillation with reward-based RL, such as SDAR (Lu et al., 2026): their students also learn from task reward, which would confound the comparison; a repaired teacher could serve as the distillation term of such a hybrid.

5.2 Main Results

R2S beats every baseline in all twelve cells of Table 2 and ends above the teacher acting alone. The gain over vanilla OPD averages +17.8 points from the 9B teacher and +9.8 from the 30B-A3B teacher. Averaged over the ten cells outside the 30B-A3B teacher's WebShop pair, the vanilla student ends within a point of the teacher acting alone. On WebShop the 30B-A3B teacher acting alone seldom completes a purchase, since most of episodes run out of turns, and every student ends above it. Figure 3 follows the Qwen3.5-9B-to-4B students of all five methods over three training runs each: the R2S student separates from every baseline well before the final checkpoint.

The teacher can be the student's own base model. Table 3 repeats the Qwen3.5-4B cells with the 4B as its own teacher. In all three environments, the R2S student taught by the repaired 4B passes the vanilla student that the 9B teaches in Table 2. Since the teacher and the student are the same

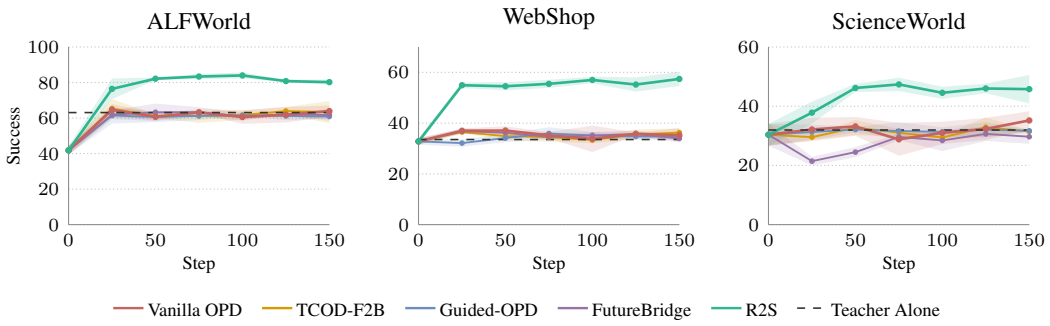

Figure 3: **R2S students separate from every baseline during training.** Qwen3.5-9B teaches Qwen3.5-4B; mean ± one standard deviation over three training runs; dashed: the teacher acting alone. Other teacher–student pairs are in Appendix F.

Table 3: **Self-distillation: Qwen3.5-4B teaches Qwen3.5-4B.** R2S first repairs the 4B teacher on its own takeovers; the other arms learn from the unrepaired 4B.

Environment	Student alone	Vanilla OPD	TCOD-F2B	Guided-OPD	FutureBridge	**R2S (ours)**
ALFWorld	43.4±1.0	49.9±2.6	49.7±2.6	48.5±3.8	50.6±1.9	72.7±1.3
WebShop	33.7±0.9	33.3±2.2	31.3±0.9	34.0±1.8	31.1±1.5	58.1±2.5
ScienceWorld	28.0±1.5	28.2±2.3	27.2±0.8	29.3±1.8	29.2±2.9	42.4±3.4

model, the student–teacher gap that ranks states is zero, so the pipeline draws one turn per episode of the untrained student at random under the same goal cap and budget. What lifts the student is not a larger model's knowledge but scores informed by the ending, which its own base model can supply.

5.3 Ablations

Each ablation below varies one part of R2S in the Qwen3.5-9B-to-4B cells of Table 2, on all three environments.

Takeover selection. Table 4 repairs the teacher with three rules for picking the winning takeovers. Ground truth reads the task reward; the consensus vote and the LLM judge do not. With either reward-free rule the student ends above vanilla OPD in all three environments, so the repair depends neither on reading the reward nor on one particular rule.

Table 4: Ablation on **takeover selection**. Success (%). Ground truth reads the task reward and is not ranked.

	ALFWorld	**WebShop**	**ScienceWorld**
Vanilla OPD	65.8±2.6	34.6±2.9	34.6±3.5
Consensus vote	81.9±2.4	63.1±1.6	39.5±2.9
LLM judge	82.6±0.9	55.5±3.0	50.3±1.5
Ground truth	85.9±1.2	65.0±1.4	49.0±1.1

Effect of informed targets. Table 5 separates the takeovers from the informed targets built on them. The label-SFT control fine-tunes the teacher on the selected takeovers by standard SFT. Its student already ends well above vanilla OPD in all three environments, so the takeovers carry signal of their own; fitting the informed targets instead lifts the student further. The takeovers help most as context for the teacher's scores of the student's own tokens, not as SFT data.

Table 5: Ablation on **components**. Success (%); the rows are vanilla OPD, label-SFT and R2S.

Takeovers	**Informed targets**	**ALFWorld**	**WebShop**	**ScienceWorld**
✗	✗	65.8±2.6	34.6±2.9	34.6±3.5
✓	✗	76.8±3.5	46.7±1.3	41.0±3.0
✓	✓	81.9±2.4	55.9±1.4	47.8±3.0

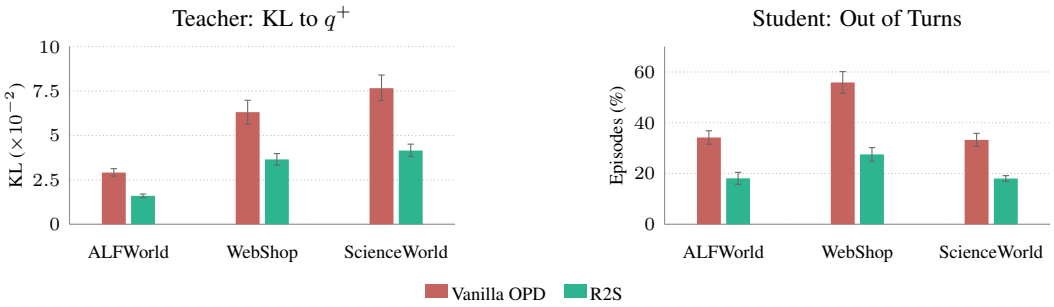

Figure 4: **Repair moves the teacher toward its informed targets, and its student runs out of turns about half as often.** Left: each teacher's top-16 KL to q^+ on held-out states, with 95% goal-bootstrap intervals. Right: held-out episodes that end unfinished at the turn budget, mean ± std over three decodes.

Table 7: **The repaired student spends fewer turns searching.** Turns per episode in the last ten training updates, mean±std over three training runs. *Task* turns act on the goal, such as taking or buying.

Method	ALFWorld			WebShop			ScienceWorld		
	Turns	Searching	Task	Turns	Searching	Task	Turns	Searching	Task
Vanilla OPD	20.7±0.9	15.9±1.2	4.0±0.3	11.5±0.4	7.8±0.5	3.6±0.1	18.2±0.5	10.9±0.6	6.9±0.7
R2S	15.6±0.5	11.4±0.6	3.7±0.3	8.8±0.2	4.1±0.1	4.7±0.1	14.6±0.3	8.2±0.6	6.1±0.3

Supervised turns. R2S fits the informed targets on the suffix of the student's episode, from the divergence turn to its end. Table 6 supervises instead the divergence turn alone or the prefix up to it. Every span lifts the student well above vanilla OPD in all three environments, so the informed targets help whichever turns they cover. Vanilla OPD, whose teacher scores each token without seeing how the episode ends, trails them all by a wide margin.

Table 6: Ablation on **supervised turns**. Success (%); the suffix, from the divergence turn to the end, is R2S.

	ALFWorld	**WebShop**	**ScienceWorld**	**Avg.**
Divergence turn	82.5±2.3	48.5±1.4	48.7±0.4	59.9
Prefix	81.0±1.8	52.7±0.8	47.6±2.1	60.5
Suffix	81.9±2.4	55.9±1.4	47.8±3.0	61.9

5.4 Long-Horizon Behavior

We compare the Qwen3.5-9B-to-4B R2S teachers and students of Table 2 with vanilla OPD and the label-SFT control of Section 5.3. Paired intervals are 95% goal-bootstrap intervals.

Repair brings the teacher closer to its informed targets, and its student runs out of turns less often. Scored against the same q^+ targets over the complete assistant suffix of held-out states, the repaired teacher is closer to them than the unrepaired teacher in every environment, as the left panel of Figure 4 shows, and closer than the label-SFT teacher, with paired intervals below zero, as Tables 18 and 19 show. In the evaluation episodes behind Table 2, the student trained on the repaired teacher runs out of turns about half as often as the vanilla OPD student, as the right panel shows.

The saved turns are searching turns. Table 7 splits the turns of the students' training rollouts into searching, that is moving, opening and looking, or searching and paging on WebShop, and acting on the task. The R2S student spends fewer turns searching than the vanilla OPD student in every environment and about as many on the task, also when only the episodes each student wins are counted. This matches the direction of the informed scores in Table 1, which fall most on the student's hedging and searching words.

Its students finish more tasks and fail on fewer turns. In a separate decoding batch described in Appendix G.2, which keeps every initial goal in the denominator, Figure 5 shows the R2S student completing more tasks than vanilla OPD and label-SFT in every environment. It also ends with

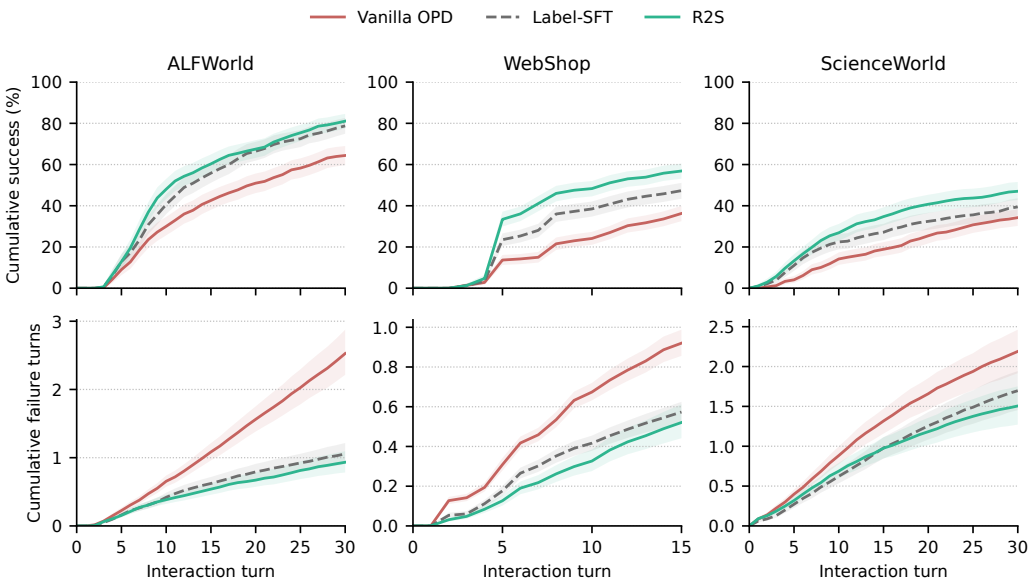

Figure 5: **R2S students finish more tasks and fail on fewer turns.** Qwen3.5-9B teaches Qwen3.5-4B. Success (%) and mean observed-failure turns (invalid or malformed actions) per episode, accumulated up to each interaction turn, from a separate decoding batch that keeps every initial goal. Bands are 95% stratified goal-bootstrap intervals.

fewer observed-failure turns than vanilla OPD everywhere, each paired interval excluding zero, and the gap grows over the episode.

6 Conclusion

The teacher's takeovers can still finish from the states where the student departs most from it, yet its plain scores let the student's delay pass; only a finishing takeover in context turns them against it. R2S transfers information from selected teacher continuations into the scores used by ordinary OPD. Its repair supervises every assistant position from the divergence turn onward with takeover-informed targets, then freezes the repaired teacher as the student's sole scorer. Under a shared protocol, R2S outperforms three state-of-the-art agentic-OPD methods in all twelve model–environment cells, and it also leads when the teacher is the student's own base model.

Limitations

Compute. Before distillation, R2S spends the untrained student's episodes, $N \times K = 19{,}200$ takeover episodes, the selection step and one teacher fine-tune. Every method then trains for 150 steps; Guided-OPD and FutureBridge also spend teacher generation inside training. **One model family.** Evidence comes from text-agent benchmarks and Qwen3 and Qwen3.5 models. **Replayable environments.** Takeovers start from the student's own states, so R2S needs an environment that can be replayed to a given state; all three benchmarks here can.

Reproducibility Statement

The distillation objective, environments, budgets, and evaluation protocol are specified in Sections 3 and 5.1 and Appendix D. The R2S hyperparameters (N, K, m, support size, epochs) and the state-selection, takeover, selection and fine-tuning procedures are in Section 4 and Appendix C. Configuration files, prompts, environment wrappers, model and tokenizer revisions, and seeds will be included with the anonymized code in the supplementary material.

AI USE STATEMENT

Large language models were used as writing and coding assistants: to restructure and draft the manuscript from the authors' experimental records and notes, to write the TikZ and pgfplots code for the figures, and to assist with analysis and plotting scripts. All experiments, their design, and their interpretation are the authors'. Every number in the text and tables was checked by an author against the frozen evaluation records, and the authors take full responsibility for the content.

ETHICS STATEMENT

This work studies training procedures for language-model agents on public text-based benchmarks (ALFWorld, WebShop, ScienceWorld) and involves no human subjects or personal data. Its methodological finding, that a capable teacher can supply harmful token-level supervision, argues for auditing distillation pipelines rather than trusting standalone scores.

REFERENCES

Chishui Chen, Yaoyou Fan, Te Sun, Yi Yang, Chenghao Sun, Delin Mao, Hongbo Qiao, Zuowei Zhang, Junxi Wang, Chenxing Sun, et al. Look Ahead Before You Distill: Future Trajectory Validation of Teacher Guidance for Agentic On-Policy Distillation. *arXiv preprint arXiv:2608.01953*, 2026.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-Group Policy Optimization for LLM Agent Training. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

John Gkountouras, Josip Jukić, and Ivan Titov. Consensus as Privileged Context for Label-Free Self-Distillation. *arXiv preprint arXiv:2607.13643*, 2026.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, et al. Reinforcement Learning via Self-Distillation. *arXiv preprint arXiv:2601.20802*, 2026.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. *arXiv preprint arXiv:2503.09516*, 2025.

Simran Kaur, Narutatsu Ri, Yinghui He, Liam Fowl, and Sanjeev Arora. Rethinking On-Policy Self-Distillation for Thinking Models. *arXiv preprint arXiv:2607.05184*, 2026.

Gengsheng Li, Mao Zheng, Mingyang Song, Ruiqi Liu, Tianyu Yang, Jie Sun, Qiyong Zhong, Haiyun Guo, Junfeng Fang, Dan Zhang, et al. On-Policy Distillation with Curriculum Turn-level Guidance for Multi-turn Agents. *arXiv preprint arXiv:2606.15912*, 2026a.

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

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

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.

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

Qwen Team. Qwen3.5: Towards Native Multimodal Agents, 2026. URL https://qwen.ai/blog?id=qwen3.5.

Stéphane Ross, Geoffrey J. Gordon, and Drew Bagnell. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300*, 2024.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-Distillation Enables Continual Learning. *arXiv preprint arXiv:2601.19897*, 2026.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew J. Hausknecht. ALFWorld: Aligning Text and Embodied Environments for Interactive Learning. In *International Conference on Learning Representations (ICLR)*, 2021.

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.

Ruoyao Wang, Peter A. Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. ScienceWorld: Is your Agent Smarter than a 5th Grader? In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Xing Jin, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, et al. RAGEN: Understanding Self-Evolution in LLM Agents via Multi-Turn Reinforcement Learning. *arXiv preprint arXiv:2504.20073*, 2025.

Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, et al. MiMo-V2-Flash Technical Report. *arXiv preprint arXiv:2601.02780*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhuolin Yang, Zihan Liu, Yang Chen, Wenliang Dai, Boxin Wang, Sheng-Chieh Lin, Chankyu Lee, Yangyi Chen, Dongfu Jiang, Jiafan He, et al. Nemotron-Cascade 2: Post-Training LLMs with Cascade RL and Multi-Domain On-Policy Distillation. *arXiv preprint arXiv:2603.19220*, 2026.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. WebShop: Towards Scalable Real-World Web Interaction with Grounded Language Agents. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-Policy Context Distillation for Language Models. *arXiv preprint arXiv:2602.12275*, 2026.

Weichen Yu, Xiaomin Li, Yizhou Zhao, Xiaoze Liu, Ruowang Zhang, Haixin Wang, Yinyi Luo, Chen Henry Wu, Gaurav Mittal, Matt Fredrikson, et al. Multi-Rollout On-Policy Distillation via Peer Successes and Failures. *arXiv preprint arXiv:2605.12652*, 2026.

Weixiao Zhan, Yongcheng Jing, Leszek Rutkowski, and Dacheng Tao. Distillation Traps and Guards: A Calibration Knob for LLM Distillability. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-Distilled Reasoner: On-Policy Self-Distillation for Large Language Models. *arXiv preprint arXiv:2601.18734*, 2026.

Yuhang Zhou, Lizhu Zhang, Yifan Wu, Mingyi Wang, Bo Peng, Jiayi Liu, Xiangjun Fan, and Zhuokai Zhao. SAGE-OPD: Selective Agent-Guided Intervention for Multi-Turn On-Policy Distillation. *arXiv preprint arXiv:2606.19659*, 2026.

Table 8: **Among these methods, only R2S repairs an external teacher without task reward and refreezes it for distillation.** RQ-free: no task-reward query for distillation supervision. The two right columns concern an external scorer, not an evolving self-teacher. †FutureBridge uses a success-selected reference pool.

Method	RQ-free	Repairs external teacher	Freeze, then distill
TCOD-F2B (Wang et al., 2026)	✓	✗	✗
Guided-OPD (Li et al., 2026a)	✓	✗	✗
FutureBridge† (Chen et al., 2026)	✓	✗	✗
ReOPD (Liao et al., 2026)	✓	✗	✗
SDFT / OPCD (Shenfeld et al., 2026; Ye et al., 2026)	✓	✗	✗
CANON (Gkountouras et al., 2026)	✓	✗	✗
SDAR (Lu et al., 2026)	✗	✗	✗
Distillation Traps and Guards (Zhan et al., 2026)	✗	✓	✓
R2S (ours)	✓	✓	✓

A Extended Related Work

A.1 Detailed Comparison with the Closest Methods

Table 8 distinguishes task-reward queries, repair of an external teacher, and freezing that repaired teacher before student distillation. Updating an EMA self-teacher during actor training is different from this separate scorer-repair stage.

Student-side curricula. TCOD-F2B (Wang et al., 2026) gradually lengthens the student's training episodes. Guided-OPD (Li et al., 2026a) assigns a decreasing share of turns to the teacher, combining cross-entropy on teacher tokens with distillation on student tokens. ReOPD (Liao et al., 2026) instead replays pre-collected teacher prefixes, reducing environment interaction while controlling the reliability of the histories scored by the teacher. These methods leave the external scorer unchanged. A repaired teacher can supply scores under such curricula as well as under ordinary OPD.

FutureBridge and teacher continuations. FutureBridge (Chen et al., 2026) is our closest experimental comparator. It continues from high-disagreement student states and validates a bridge with short student rollouts. Accepted teacher text enters the student's training buffer, alongside a curriculum over success-selected reference prefixes. The online acceptance rule reads no task reward, although the reference pool uses success selection. R2S instead uses a selected takeover as privileged context for teacher repair: the teacher learns to reproduce its informed scores without that context and is then frozen. The downstream student retains ordinary OPD.

Privileged context and consensus. SDFT (Shenfeld et al., 2026) and Self-Distilled Reasoner (Zhao et al., 2026) train an acting model from a demonstration-conditioned self-teacher. OPCD (Ye et al., 2026) also supports an external context-conditioned teacher, but transfers the context into the student directly. CANON (Gkountouras et al., 2026) supplies a label-free reference by selecting a solution with the majority answer. Our consensus selection follows this principle, applied to an external teacher's continuations from student states; the model being fitted becomes the scorer for a subsequent distillation stage.

Feedback and reward-based learning. SDPO (Hübotter et al., 2026) turns rich environment feedback into self-distillation targets; in scalar-feedback settings it uses successful attempts as feedback. SDAR (Lu et al., 2026) is a multi-turn hybrid that keeps reward-based RL as its primary objective and adds gated self-distillation. We exclude this additional student reward signal from the experimental comparison; R2S could supply the distillation scorer in a hybrid system.

External teacher calibration. Distillation Traps and Guards (Zhan et al., 2026) changes the teacher before distillation, using reinforcement fine-tuning with utility and calibration rewards and a KL anchor. R2S learns from takeover-informed soft targets on student histories through the cross-entropy of Eq. (3), without task reward, before freezing the repaired teacher.

Table 9: **The states behind Table 1.** *All*: the student's highest-disagreement states with a replayable takeover; *Lost*: states at turn five or later that the student went on to lose; *Out of turns*: takeovers that reach the turn budget unfinished.

	States			Turn		Takeovers
Environment	All	Turn ≥ 5	Lost	Mean	Budget	Out of turns (%)
ALFWorld	2,399	814	489	12.4	30	55.8
WebShop	2,398	689	647	9.1	15	84.4
ScienceWorld	2,205	682	555	11.7	30	52.2

A.2 Further Related Work

OPD at scale. MiMo-V2-Flash integrates domain-specialized teachers with dense token-level supervision (Xiao et al., 2026). Nemotron-Cascade 2 uses multi-domain OPD between Cascade RL stages to recover capabilities from the strongest intermediate domain checkpoints (Yang et al., 2026). These pipelines establish the practical role of OPD in capability transfer. Our question concerns the quality of the next-token supervision supplied by a given teacher on multi-turn student histories.

Local supervision and privileged information. Li et al. (2026c) analyze teacher–student compatibility and alignment on high-probability tokens. Kaur et al. (2026) show that privileged context can harm thinking-model self-correction, motivating direct evaluation of the supervision it induces. Multi-Rollout OPD (Yu et al., 2026) enriches teacher context with peer attempts labelled by success or failure. Our references are teacher takeovers from the student's own state, selected without task reward and internalized into an external scorer. We assess that scorer through target fidelity and the completion behavior of its distilled students.

Reinforcement learning for multi-turn agents. OPD teachers are often RL-trained. PPO constrains updates with a clipped ratio and a learned value function, GRPO estimates relative advantages from groups of rollouts, and Search-R1 applies group-relative RL to interleaved reasoning and retrieval (Schulman et al., 2017; Shao et al., 2024; Jin et al., 2025). Sparse terminal returns make trajectory-level credit coarse in longer interactions. RAGEN documents recurrent instability, and Turn-PPO and GiGPO estimate turn- and step-level advantages (Wang et al., 2025; Li et al., 2026b; Feng et al., 2025). These methods use task reward to improve the acting policy. Our work begins after RL. It shows that a teacher's token scores can let through delays that outcome reward would punish, and it repairs those scores without querying reward.

B Analysis Details

The table of Section 3.2. Table 1 leaves out the states before turn five, where a takeover is close to the teacher acting alone from the start; each state has $K{=}8$ takeovers, and *rescued* is taken over the states the student went on to lose. Success is the environment's own outcome. The drop columns of Table 1 keep those of these states where at least one takeover finishes and read the untrained student's turns up to and including the mined one. The teacher scores the student's own tokens twice, plain and with that takeover in context, as described in Appendix C; a drop entry is $100\,(1 - e^{\bar{d}})$, where $\bar{d}$ is the mean change in log-probability over the student's tokens of that kind. Hedging and searching words are thinking tokens from the two lists below; *other* is the rest of the thinking. The out-of-turns shares that Section 3.2 cites, shown in the right panel of Figure 4, leave these states: they give the share of held-out evaluation episodes that reach the turn budget unfinished for the 9B-taught 4B students of Table 2 at their final checkpoint, in the three decodes behind their success there, trained with vanilla OPD on the plain scores or with R2S on the repaired teacher (seen and unseen tasks averaged on ALFWorld and ScienceWorld, the 500 canonical goals on WebShop); Table 11 adds the untrained student, the teacher acting alone and the other endings, on the seen split. Table 9 gives the states and Table 10 the scores, with intervals, the student's words of committing (*take, put, use, buy, now, done*, ...), which fall less than its hedging words and more than the rest, and its action tokens, which barely move. Hedging words are *maybe, might, likely, probably, perhaps, could, possibly, seem(s), guess, assume, unsure, potentially, whether, if, unless, unclear, unknown, hopefully, presumably*; searching words are *look, search, check, examine, explore, find, locate, scan, browse* with their inflections, and *inventory*.

Table 10: **The teacher's scores behind the drop columns of Table 1.** *Scored*: states where at least one takeover finishes, and the student's turns through the mined one; *Against other*: a class's mean change in log-probability over that of the rest of the thinking, with a 95% interval over states.

	Scored		Against other		Drop (%)	
Environment	States	Turns	Hedging	Searching	Committing	Action
ALFWorld	458	4,957	2.4 [2.1, 2.8]	1.8 [1.4, 2.2]	6.1	0.1
WebShop	107	906	2.0 [1.3, 2.9]	1.9 [0.9, 3.0]	6.8	0.5
ScienceWorld	403	4,085	2.7 [2.4, 3.1]	3.5 [3.0, 4.1]	11.3	0.3

Table 11: **How the episodes end.** Evaluation episodes of the 9B teacher acting alone and of its 4B students, on the seen split (WebShop: the canonical goals); mean±std over three decodes. *Turns*: mean turns per episode.

Environment	Policy	Success (%)	Ran out of turns (%)	Ended otherwise (%)	Turns
ALFWorld	teacher alone	60.2 ± 1.5	39.8 ± 1.5	0.0	20.0
	student, untrained	45.0 ± 3.6	55.0 ± 3.6	0.0	22.3
	vanilla OPD student	65.7 ± 4.7	34.3 ± 4.7	0.0	19.0
	R2S student	80.5 ± 2.2	19.5 ± 2.2	0.0	14.7
WebShop	teacher alone	33.5 ± 0.3	56.7 ± 0.9	9.8	11.9
	student, untrained	33.7 ± 0.9	56.7 ± 1.2	9.6	11.9
	vanilla OPD student	34.6 ± 2.9	55.9 ± 4.3	9.5	11.8
	R2S student	55.9 ± 1.4	27.5 ± 2.7	16.6	9.0
ScienceWorld	teacher alone	33.6 ± 3.6	35.7 ± 1.9	30.7	19.7
	student, untrained	26.7 ± 1.8	38.6 ± 2.1	34.8	17.7
	vanilla OPD student	38.8 ± 2.2	31.9 ± 1.5	29.3	18.9
	R2S student	52.6 ± 3.5	17.1 ± 0.7	30.2	14.5

Running out of turns is how the teacher and the vanilla student fail. Table 11 sorts the evaluation episodes of the 9B teacher and of its 4B students by how they end. The turn budget is 30 turns on ALFWorld and ScienceWorld and 15 on WebShop; an episode that ends otherwise is, on WebShop, a purchase that was not the target and, on ScienceWorld, a task that ended without completion. On ALFWorld every failure is a timeout, because the wrapper ends an episode only on success; on WebShop more than half of the teacher's and of the vanilla student's episodes run out the fifteen turns without a purchase. In every environment the vanilla student takes about the teacher's number of turns and times out almost as often. The R2S student finishes in fewer turns and times out far less often; on WebShop it also buys more often. At the mined states themselves the teacher's takeovers win more often than the student's own continuation, as shown in Table 13; what removes the delay is the takeover-informed scoring that the repair distills into the teacher.

The R2S student spends fewer turns searching. Table 12 gives the full readout from the students' own training rollouts for one training run per method, and Table 7 averages its searching and task turns over three runs. *Searching* counts the turns spent moving, opening and looking (ALFWorld: go to, open, close, look, examine, inventory; ScienceWorld: teleport, look, read, wait, open, close; WebShop: search, next or previous page, back to search); *task* counts the turns spent acting on the goal (ALFWorld: take, move, put, clean, heat, cool, use; ScienceWorld: focus, pick up, move, pour, use, activate, connect and the like; WebShop: opening an item, choosing an option, reading its description, buying). *Revisits* are moves to a receptacle or room already visited in the episode, as a share of all moves; *hedges* are maybe, might, likely, perhaps, if, whether and the like. Over training the vanilla student's episodes get shorter, not longer, to about the length of the teacher's. In every environment the R2S student's saving is in the turns spent searching. Counting only the episodes each student wins, the R2S student still spends fewer turns searching, averaged over the three runs, 8.4 against 10.5 on ALFWorld, 2.0 against 3.0 on WebShop and 7.1 against 9.1 on ScienceWorld, and about as many on the task. On ALFWorld it visits fewer receptacles and opens fewer of them, at about the same rate of revisits, so it finds the object sooner rather than wandering less; on ScienceWorld it spends fewer turns looking, reading and waiting, a share of its turns that grows in the vanilla student as training goes on; on WebShop it searches, pages and backs out less, and opens items, picks options and buys more.

Table 12: **What the trained students do.** Training rollouts of the 9B→4B students: *untrained* is the first ten updates of the vanilla run, the other rows the last ten updates of each run. Turns, searching and task turns are per episode; hedges per hundred thinking words, tokens per turn.

					Turns per episode			Thinking	
Environment	Student	Success (%)	Turns	Ran out (%)	searching	task	Revisits (%)	hedges	tokens
ALFWorld	untrained	45.6	22.8	54.4	18.5	3.7	11	1.81	28
	vanilla OPD student	64.4	20.1	35.6	15.4	3.8	13	1.89	33
	R2S student	79.4	15.3	20.6	11.6	3.4	11	1.24	28
WebShop	untrained	24.6	12.8	67.5	9.2	3.6	–	1.31	59
	vanilla OPD student	32.3	11.4	50.9	7.7	3.6	–	1.46	63
	R2S student	50.9	8.7	24.2	4.0	4.6	–	1.42	75
ScienceWorld	untrained	26.9	20.5	48.1	12.4	7.8	57	1.80	41
	vanilla OPD student	35.6	18.4	30.6	11.6	6.4	46	1.90	43
	R2S student	58.1	14.3	16.9	7.8	6.1	46	1.75	51

C R2S Implementation Details

Untrained-student episodes. Before training, the untrained student plays the training-split tasks under the training prompt at T=1.0, for the episode budget of a training run (150 steps); evaluation tasks are discarded before scoring.

Mining score. In Δ_t of Eq. (2), π_θ is the untrained student, and each term is the sampled per-token estimate of the reverse KL of Eq. (1), so a single token's term can take either sign. Action tokens stay out of Δ_t because the thinking before them decides the action; with a finishing takeover in context, the teacher's scores of the student's action tokens barely move, as Table 10 shows.

Filters and mining. A turn is eligible when its thinking span has at least three tokens and when it ends in a tool call. Within each of these episodes we keep the eligible turn with the largest Δ_t of Eq. (2) (one state per episode), cap three states per goal, and keep the top N=2,400 by Δ_t.

Takeovers. From each restored state the teacher attempts the episode K=8 times independently. Takeovers are sampled at T=1.0 with reasoning mode off, up to the environment's turn budget.

Selection. As described in Section 4, any check that tells a finishing takeover from a failing one can select the takeovers. We use two reward-free checks. The consensus vote groups a state's takeovers by how they end and keeps the state when the largest group holds at least m=2 of them; the LLM judge, Qwen3.5-9B with reasoning mode on, reads each takeover and keeps the state when it approves at least m=1. Either way the shortest passing takeover is selected. Apart from the ground-truth row of Table 4, task reward is recorded for accounting only and is never read by mining or selection.

Privileged targets and diagnostic plain scores. The frozen teacher T_0 scores the student's unmodified token stream through the end of the episode. In the privileged pass the selected takeover is inserted as one extra user message immediately before the student's first turn. The message is headed "Reference walkthrough" and lists the teacher's thinking and action for each takeover step; only the teacher's steps are shown, not the replayed prefix. The header describes the walkthrough as a successful expert completion of the same task, although the selected takeover need not have succeeded. Every position of the student stream shifts by one constant offset, so the two passes align position by position. The plain pass scores the same stream without the message (closed-book). Both passes cache per-position top-16 (token, log-probability) lists. R2S trains only on the privileged targets; the plain pass is the reference of the drop columns of Table 1.

Positions and targets. In R2S, every assistant position from the mined turn, inclusive, to the original student's final assistant turn receives q^+. Assistant turns without a tool call are included. Earlier history remains causal context; prompts and environment observations are not targets. The cached reference lists teacher thinking and actions without future environment observations. It informs targets on the original student text and is not substituted as training output.

Drop rules. A state is dropped when its prompt cannot be re-tokenized identically, when the privileged stream exceeds 32,768 tokens, or when it has no privileged position.

Table 13: **Takeovers in every cell of Table 2.** Success of $K=8$ takeovers from each of the untrained student's 2,400 highest-disagreement states, beside the student's own episodes and *teacher alone*, the teacher acting from the first turn as in Table 2. The takeovers beat the untrained student's own outcome in eleven of twelve cells.

Student	Environment	States	Student (%)	Takeover (%)	States with a win (%)	Teacher alone (%)
Teacher Qwen3.5-9B						
Qwen3.5-2B	ALFWorld	2,391	11.9	54.6	83.4	63.1
Qwen3.5-4B	ALFWorld	2,399	51.8	56.7	82.0	63.1
Qwen3.5-2B	WebShop	2,399	5.0	21.8	40.8	33.5
Qwen3.5-4B	WebShop	2,398	19.6	22.0	39.0	33.5
Qwen3.5-2B	ScienceWorld	2,272	1.2	31.2	70.6	32.0
Qwen3.5-4B	ScienceWorld	2,205	22.4	35.6	73.4	32.0
Teacher Qwen3-30B-A3B-Instruct-2507						
Qwen3-4B	ALFWorld	2,400	33.3	41.8	58.4	42.3
Qwen3-8B	ALFWorld	2,400	45.5	42.2	56.5	42.3
Qwen3-4B	WebShop	2,400	10.4	14.0	28.2	2.9
Qwen3-8B	WebShop	2,400	20.1	31.7	47.2	2.9
Qwen3-4B	ScienceWorld	2,259	10.2	36.6	65.9	32.9
Qwen3-8B	ScienceWorld	2,159	21.6	38.8	66.5	32.9

Loss and fine-tuning. The loss at each labeled position is the cross-entropy between the target q^+, kept on its top-16 tokens and renormalized, and the model's full-vocabulary log-softmax. The targets are computed once from the frozen T_0 and are not recomputed as T_ϕ trains. As in Eq. (3), we average all selected assistant positions within each state, then average states. The teacher is initialized from T_0 and fine-tuned full-parameter for one epoch at learning rate 2×10^{-6}. The optimizer is AdamW with betas $0.9/0.95$ and no weight decay, with 10 warmup steps followed by cosine decay to $0.1\times$ and gradient clip 1.0. Batches hold 8 labeled states, 10% of goals form a goal-disjoint held-out split, and training runs in bf16. For Qwen3.5-9B the vision tower stays frozen, so full-parameter means the text decoder.

Downstream distillation. The repaired teacher is refrozen and scores otherwise ordinary OPD in place of T_0, with the sampled estimator, or with the top-16 estimator on the teacher's own top-16 support in Table 17. The original teacher takes no part in the student's training.

Takeover pools. Every cell of Table 2 mines 2,400 states from the episodes of its own untrained student and replays them; Table 13 gives the takeover statistics of every cell. In Table 13, *States* counts the mined states with at least one replayable takeover, *Student* is the untrained student's own success on the same episodes, and *teacher alone* is the teacher acting from the first turn (mean of its evaluation splits); success is the environment's own outcome (ALFWorld won, WebShop exact success, ScienceWorld completion). The takeovers beat the untrained student's own outcome in eleven of twelve cells.

Self-distillation. In Tables 3 and 15, teacher and untrained student are the same Qwen3.5-4B checkpoint, so Δ_t of Eq. (2) is zero up to numerical noise and cannot rank states. The self pipeline instead draws one eligible turn per episode of the untrained student at random, under the same goal cap and budget, and the 4B itself runs the takeovers and the selection, so no larger model enters the pipeline. Repair and distillation follow the settings above. The vanilla control, TCOD-F2B, Guided-OPD and FutureBridge learn from the unrepaired 4B teacher. The repaired 4B acting alone, T_ϕ alone in Table 15, passes the unrepaired 9B acting alone on every split. The Student alone column is the untrained 4B of Tables 2 and 14.

D Experimental Protocol

Environments. ALFWorld, WebShop and ScienceWorld are the three environments. WebShop (Yao et al., 2022) reports success on the 500 canonical evaluation goals and on a disjoint 256-goal pool, with 15 steps per episode. ALFWorld (Shridhar et al., 2021) reports seen and unseen success on 140 and 134 held-out games under a TCOD-aligned full-history protocol with 30 environment steps. Our wrapper ends an episode only on success. ScienceWorld (Wang et al., 2022) follows the TCOD splits. The first half of each of the 17 training task types' variations trains, 140 variations sampled

from the second half form the seen evaluation, and the first 134 rows of the frozen 1,308-task test set of unseen types form the unseen evaluation. Episodes run 30 steps under the same full-history tool-call protocol. The metric is success, the simulator's completion flag, as on ALFWorld and WebShop; TCOD reports the simulator's dense task score, which Table 16 gives for every ScienceWorld cell of the main line. The simulator and the action and object lists shown to the agent are TCOD's.

Models. The main line distills Qwen3.5-9B (Qwen Team, 2026) into Qwen3.5-4B and Qwen3.5-2B, and Qwen3-30B-A3B-Instruct-2507 (Yang et al., 2025) into Qwen3-4B and Qwen3-8B; Qwen3.5-4B also teaches itself in Table 3. Model and tokenizer revisions, prompts, environment wrappers and checkpoint provenance will be included in the released configuration files.

Prompt protocol. Every arm and anchor of Table 2 uses one prompt per environment that carries one fixed worked demonstration, rendered in each model's native tool-call template; native thinking is off, and the prompt asks for one to three sentences of plain-text deliberation before exactly one tool call per turn, a deliberation that stays part of the generated response.

Objective. The vanilla OPD, R2S and TCOD-F2B rows of Table 2 minimize the reverse KL of Eq. (1) with the sampled estimator of Eq. (4), consumed as a per-token advantage with coefficient β=1. Guided-OPD selects the teacher or the student to act at each turn and trains the student with cross-entropy on teacher-generated tokens and the same sampled reverse KL on student-generated tokens; FutureBridge applies the same sampled reverse KL to episodes that carry an accepted teacher bridge. The top-16 estimator of Eq. (1) is the control of Table 17. Task reward is logged and zeroed for every distillation arm. The sampled estimator gives each student token the advantage (Xiao et al., 2026; Li et al., 2026c),

$$A_{t,i} = -\beta[\log \pi_\theta(y_{t,i} \mid h_{t,i}) - \log T(y_{t,i} \mid h_{t,i})], \qquad \tau \sim \pi_\theta, \tag{4}$$

with observations loss-masked and updates restricted to student tokens.

Training. Every distillation run takes 150 optimization steps and evaluates every 25 steps. Each step trains on 16 episodes, 128 on WebShop (16 goals $\times$ 8 samples). The rollout temperature is 0.4. The optimizer is Adam at a constant learning rate of 10^{-6} with betas $(0.9, 0.98)$, weight decay 0.1 and gradient clip 1.0. In every sampled-KL arm the distillation log-ratio of Eq. (4) enters the clipped PPO surrogate (ratio clip 0.2) as the per-token advantage; in the top-16 twins of Table 17 every advantage is zero and the clip is inert. There is no additional KL or entropy term. The per-turn cap is 512 tokens in training and 4,096 at evaluation. Student responses are copied byte-identically into the separately served teacher context. Reasoning mode is off for acting, scoring and distillation.

Baselines. TCOD-F2B, Guided-OPD and FutureBridge share this protocol and keep their own method settings.

Evaluation. Each main-line cell is the final checkpoint of one training run, decoded three times on the evaluation splits above: its in-training evaluation and two re-loads under further decoding seeds. Entries are the mean$\pm$std over the three decodes, and Figure 3 adds three training runs per method for the Qwen3.5-9B-to-4B pair. Student-alone anchors are three decodes of the untrained student under the same prompt. An evaluation in which an episode returned no outcome after an infrastructure failure is repeated at unchanged weights and decoding seed.

Teacher acting alone. The teacher itself is evaluated as the acting policy through the same evaluation path, prompts and decoding seeds (T=0.4); its split means are the teacher-alone anchors of Table 2 and Table 13, and Table 14 gives them per split. On WebShop the 30B-A3B teacher seldom completes a purchase: 94% of its episodes end at the turn limit, and on 13.7 of its turns per episode on average it answers without calling the action tool.

E Additional Results

E.1 Every Split of the Main Results

The main text averages the splits. Each cell of Table 2 and Table 3 is, per decoding seed, the mean of the seen and unseen values of the same evaluation (ALFWorld, ScienceWorld) or the success rate on the 500 canonical WebShop goals; the mean and std are then taken over the three decoding seeds. Table 14 and Table 15 give every split, and for WebShop the disjoint 256-goal pool, a diagnostic set. R2S exceeds every baseline in mean on all 24 splits.

Table 14: **Main results by split.** The cells of Table 2 before averaging: ALFWorld and ScienceWorld seen / unseen, WebShop the 500 canonical goals / the disjoint 256-goal pool; protocol as in Table 2. Bold: the best of the five trained arms where its gap to the runner-up exceeds the larger decode std; ties are unmarked.

Student	Split	Student alone	Vanilla OPD	TCOD-F2B	Guided-OPD	FutureBridge	R2S (ours)
Panel A: Qwen3.5-9B → Qwen3.5-2B / 4B, three environments							
ALFWorld (success, %)						*teacher alone: seen* 60.2±1.5, *unseen* 65.9±3.0	
Qwen3.5-2B	seen	19.8±1.5	55.0±2.6	59.0±2.2	60.2±4.8	54.0±2.9	**76.9±3.2**
	unseen	10.4±3.7	60.9±2.3	66.4±2.0	66.9±1.1	59.5±4.3	**78.9±3.0**
Qwen3.5-4B	seen	45.0±3.6	65.7±4.7	59.3±2.6	60.2±4.4	58.1±2.9	**80.5±2.2**
	unseen	41.8±3.9	65.9±1.1	60.0±1.6	62.9±1.1	58.7±4.5	**83.3±3.1**
WebShop (success, %; 500 canonical goals / 256-goal pool)						*teacher alone: canonical* 33.5±0.3, *pool* 28.6±2.4	
Qwen3.5-2B	canonical	17.6±1.1	35.7±1.6	33.1±0.5	35.6±0.6	36.3±1.0	**57.9±1.6**
	pool	16.0±4.1	33.5±0.2	29.9±1.6	32.0±1.7	31.9±2.5	**52.2±1.8**
Qwen3.5-4B	canonical	33.7±0.9	34.6±2.9	35.1±0.9	35.5±1.3	33.9±1.7	**55.9±1.4**
	pool	31.5±1.6	32.8±1.8	32.7±0.6	31.4±1.9	30.5±1.4	**54.7±1.7**
ScienceWorld (success, %)						*teacher alone: seen* 33.6±3.6, *unseen* 30.3±0.4	
Qwen3.5-2B	seen	2.4±1.1	25.5±1.1	34.0±2.5	31.2±4.9	18.1±5.5	**41.7±1.5**
	unseen	3.7±1.3	21.9±4.1	21.9±3.4	20.6±2.6	19.9±1.7	**34.1±0.4**
Qwen3.5-4B	seen	26.7±1.8	38.8±2.2	32.1±0.7	35.0±3.7	32.4±0.8	**52.6±3.5**
	unseen	29.4±1.1	30.3±5.0	29.1±1.3	29.4±1.6	29.1±2.0	**43.0±3.0**
Panel B: Qwen3-30B-A3B-Instruct-2507 → Qwen3-4B / 8B, three environments							
ALFWorld (success, %)						*teacher alone: seen* 43.1±2.2, *unseen* 41.5±2.6	
Qwen3-4B	seen	27.4±0.4	41.9±0.8	41.7±1.1	42.9±1.4	44.5±0.4	**54.8±1.1**
	unseen	33.3±4.2	40.5±2.8	43.5±1.1	41.0±1.3	39.1±5.6	**54.2±1.9**
Qwen3-8B	seen	45.2±1.1	46.0±0.4	48.6±1.9	47.1±1.9	45.2±2.2	**56.4±0.7**
	unseen	48.5±0.7	43.0±4.8	49.0±2.4	46.8±5.1	46.5±2.8	**53.0±1.3**
WebShop (success, %; 500 canonical goals / 256-goal pool)						*teacher alone: canonical* 2.9±0.3, *pool* 3.0±1.1	
Qwen3-4B	canonical	15.5±1.2	24.7±1.1	22.8±0.7	23.3±1.6	16.7±0.9	**28.5±0.6**
	pool	13.8±0.5	21.6±1.5	21.1±2.0	19.1±0.4	14.5±1.4	24.6±3.5
Qwen3-8B	canonical	23.6±1.2	20.3±0.9	23.4±0.7	22.3±0.8	20.7±0.7	**30.4±1.2**
	pool	23.7±2.1	19.7±0.8	19.8±1.6	18.0±2.3	18.0±3.1	**27.9±3.2**
ScienceWorld (success, %)						*teacher alone: seen* 29.5±3.2, *unseen* 36.3±3.5	
Qwen3-4B	seen	6.9±2.7	35.2±4.3	34.3±2.5	32.6±4.2	33.1±1.8	**48.3±1.1**
	unseen	11.2±0.0	40.8±2.4	42.5±0.7	36.8±2.8	39.1±2.4	43.3±2.0
Qwen3-8B	seen	18.6±3.6	37.6±2.1	32.1±2.9	31.7±3.7	37.4±3.4	**56.2±0.4**
	unseen	29.4±5.2	36.1±2.4	37.6±2.4	34.3±0.7	36.8±3.4	**44.8±2.6**

Table 15: **Self-distillation, every split.** The cells of Table 3 by split (ALFWorld and ScienceWorld seen / unseen; WebShop canonical goals), plus T_ϕ alone: the repaired 4B, the frozen teacher of the R2S column, acting without any student training. Protocol as in Table 3. Bold: the best of the five trained arms where its gap to the runner-up exceeds the larger decode std.

Environment	Split	Student alone	Vanilla OPD	TCOD-F2B	Guided-OPD	FutureBridge	T_ϕ alone	R2S (ours)
ALFWorld	seen	45.0±3.6	53.1±2.3	50.5±3.2	50.0±4.5	56.4±3.8	71.0±2.7	**69.5±2.3**
	unseen	41.8±3.9	46.8±4.1	49.0±4.1	47.0±3.4	44.8±0.0	73.4±1.1	**75.9±1.6**
WebShop	canonical	33.7±0.9	33.3±2.2	31.3±0.9	34.0±1.8	31.1±1.5	51.7±2.8	**58.1±2.5**
ScienceWorld	seen	26.7±1.8	25.5±1.8	25.7±2.1	30.0±0.0	28.1±3.9	47.6±2.2	**45.2±5.7**
	unseen	29.4±1.1	30.8±4.6	28.6±1.1	28.6±3.5	30.3±2.8	41.5±0.9	**39.6±2.7**

E.2 ScienceWorld under the Dense Task Score

ScienceWorld dense score. The simulator also returns a dense task score with partial credit for completed subgoals, the metric TCOD reports. Table 16 gives it for every ScienceWorld cell and anchor of Table 2. Partial credit compresses the success differences; R2S leads on every seen split.

E.3 The Loss Estimator Does Not Change the Picture

The repaired teacher works under either estimator. The main line uses the sampled estimator of Eq. (4), consumed as a per-token advantage. Table 17 repeats every vanilla and R2S cell of Table 2 with the top-16 estimator of Eq. (1). For vanilla OPD the two estimators end level on average. R2S ends above vanilla OPD in all twelve cells with the sampled estimator and in eleven with the top-16 estimator. The repaired teacher is an ordinary next-token scorer whose distribution either estimator can read.

Table 16: **ScienceWorld under TCOD's dense task score.** The ScienceWorld cells and anchors of Table 2, same runs and decodes, scored by the simulator's dense task score (best score reached, %) instead of success. Bold as in Table 14.

Student	Split	Student alone	Vanilla OPD	TCOD-F2B	Guided-OPD	FutureBridge	R2S (ours)
Qwen3.5-9B teacher							
ScienceWorld (dense task score, %)					*teacher alone: seen* 58.1±3.3, *unseen* 57.0±1.0		
Qwen3.5-2B	seen	14.1±1.6	54.2±1.3	58.1±1.8	59.4±4.0	45.1±3.4	**64.7±0.8**
	unseen	21.6±2.3	50.3±2.5	50.3±2.6	53.2±2.4	48.4±1.0	**59.8±0.6**
Qwen3.5-4B	seen	50.5±3.7	62.7±1.4	57.8±1.6	59.3±4.6	58.4±0.3	**70.0±2.7**
	unseen	52.9±1.6	57.4±1.6	54.3±0.8	54.6±0.6	55.9±1.3	**64.2±0.7**
Qwen3-30B-A3B-Instruct-2507 teacher							
ScienceWorld (dense task score, %)					*teacher alone: seen* 53.7±1.4, *unseen* 57.3±1.6		
Qwen3-4B	seen	21.3±1.9	58.0±2.6	60.3±2.4	55.5±2.9	56.4±0.7	**66.5±0.9**
	unseen	30.6±1.0	63.7±1.4	64.9±0.2	60.5±1.5	62.1±2.2	63.7±1.6
Qwen3-8B	seen	44.9±2.7	60.3±3.2	56.7±3.7	53.0±0.8	61.8±3.5	**72.0±0.4**
	unseen	52.7±3.7	62.0±0.8	60.9±1.7	57.9±1.0	62.8±2.7	**66.1±1.2**

Table 17: **The loss estimator.** Vanilla OPD and R2S on the pairs of Table 2 with the sampled KL of Eq. (4) or the top-16 KL of Eq. (1); one run per cell, mean±std over three decodes of the final checkpoint. Bold: the better arm per estimator where the gap exceeds the larger decode std.

		Vanilla OPD		R2S (ours)	
Student	Split	sampled KL	top-16 KL	sampled KL	top-16 KL
Panel A: Qwen3.5-9B teacher					
ALFWorld (success, %)					
Qwen3.5-2B	seen	55.0±2.6	63.6±1.9	**76.9±3.2**	**71.9±1.8**
	unseen	60.9±2.3	66.7±6.3	**78.9±3.0**	**77.6±2.0**
Qwen3.5-4B	seen	65.7±4.7	59.8±1.8	**80.5±2.2**	**81.9±1.5**
	unseen	65.9±1.1	64.4±2.3	**83.3±3.1**	**83.6±2.0**
WebShop (success, %; 500 canonical goals / 256-goal pool)					
Qwen3.5-2B	canonical	35.7±1.6	36.4±1.1	**57.9±1.6**	**55.1±0.3**
	pool	33.5±0.2	31.2±0.7	**52.2±1.8**	**51.4±0.8**
Qwen3.5-4B	canonical	34.6±2.9	34.6±1.1	**55.9±1.4**	**49.6±1.6**
	pool	32.8±1.8	30.9±3.1	**54.7±1.7**	**47.9±2.4**
ScienceWorld (success, %)					
Qwen3.5-2B	seen	25.5±1.1	31.7±3.0	**41.7±1.5**	**48.1±4.5**
	unseen	21.9±4.1	23.1±1.3	**34.1±0.4**	**31.8±4.6**
Qwen3.5-4B	seen	38.8±2.2	32.4±7.0	**52.6±3.5**	**45.2±5.8**
	unseen	30.3±5.0	29.9±3.4	**43.0±3.0**	**39.8±6.7**
Panel B: Qwen3-30B-A3B-Instruct-2507 teacher					
ALFWorld (success, %)					
Qwen3-4B	seen	41.9±0.8	45.0±1.2	**54.8±1.1**	**54.8±1.8**
	unseen	40.5±2.8	43.8±2.3	**54.2±1.9**	**55.5±2.2**
Qwen3-8B	seen	46.0±0.4	46.2±0.4	**56.4±0.7**	**53.1±1.5**
	unseen	43.0±4.8	44.8±3.0	**53.0±1.3**	**51.7±3.0**
WebShop (success, %; 500 canonical goals / 256-goal pool)					
Qwen3-4B	canonical	24.7±1.1	**24.3±1.6**	**28.5±0.6**	17.6±2.0
	pool	21.6±1.5	**22.5±2.3**	24.6±3.5	14.5±3.3
Qwen3-8B	canonical	20.3±0.9	22.0±1.2	**30.4±1.2**	**24.3±1.4**
	pool	19.7±0.8	17.8±0.2	**27.9±3.2**	**23.6±1.2**
ScienceWorld (success, %)					
Qwen3-4B	seen	35.2±4.3	34.5±5.4	**48.3±1.1**	**46.4±4.5**
	unseen	40.8±2.4	38.1±1.5	**43.3±2.0**	39.1±4.4
Qwen3-8B	seen	37.6±2.1	28.8±2.3	**56.2±0.4**	**52.1±1.4**
	unseen	36.1±2.4	37.6±3.8	**44.8±2.6**	**43.8±5.0**

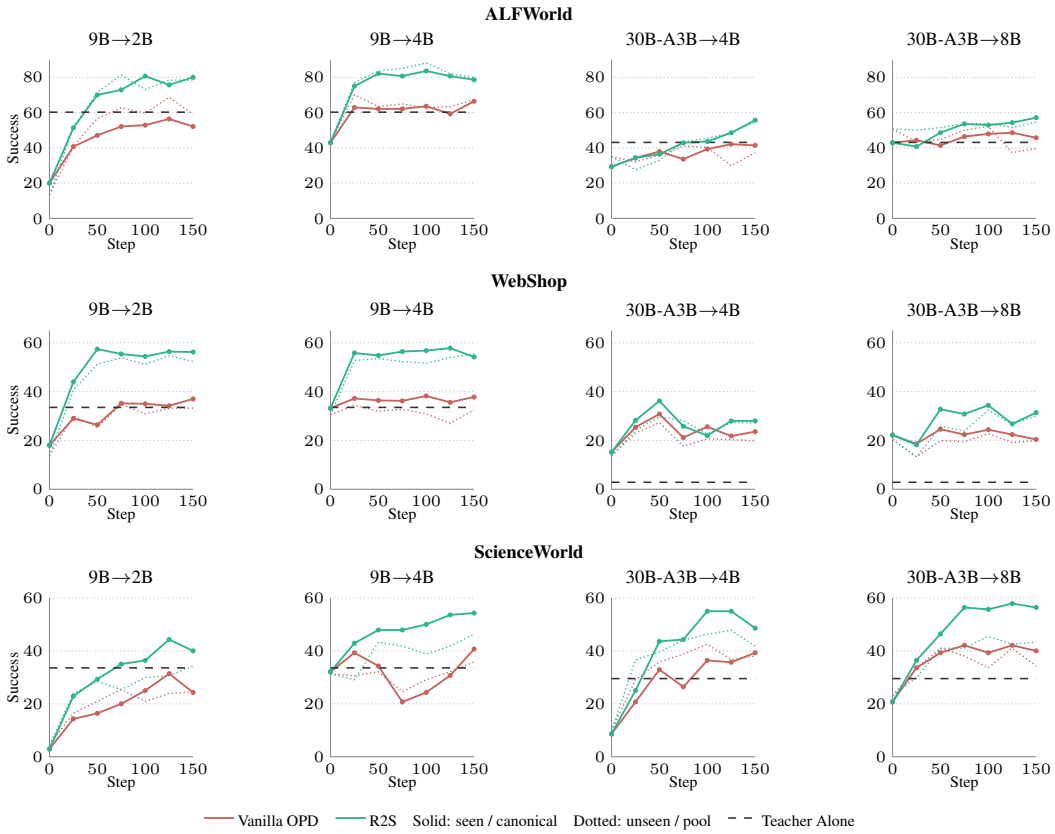

Figure 6: **Success over training in every cell of Table 2.** Vanilla and R2S students, one decode every 25 updates (update 0: untrained). Solid: seen split or WebShop's canonical goals; dotted: unseen split or its 256-goal pool; dashed: the teacher acting alone on the seen split.

F Training Curves

F.1 Every Cell of the Main Results

Training curves complement the final-checkpoint comparison. Figure 6 plots the in-training evaluation of the vanilla and R2S students of every cell of Table 2: one decode every 25 updates under the training seed, so the final point of each curve is the first of the three decodes behind the table. The R2S student ends above vanilla OPD in all twelve cells, on both splits, and above the teacher acting alone on the seen split in all twelve; no intermediate checkpoint is selected for the reported result. Figure 7 plots the sampled reverse KL to each run's own scorer. The loss falls over training in every run (mean of the last ten updates against the first ten). The original and repaired teachers define different target distributions, so the loss levels are not a common measure of task performance.

F.2 Three Training Runs per Method

Models and settings. Figure 3 follows five methods from a Qwen3.5-9B teacher to a Qwen3.5-4B student: vanilla OPD with the sampled estimator, TCOD-F2B, Guided-OPD, FutureBridge and R2S. Prompts, evaluation splits, turn budgets and optimizer settings are those of Appendix D; sampling temperature is 0.4 during both training and evaluation, and the context limit is 32,768 tokens. As in Table 2, the ALFWorld and ScienceWorld panels average the seen and unseen splits, and WebShop uses its 500 canonical goals; at every evaluation checkpoint, each run generates one episode per evaluation task.

Independent training runs. Each method has three training runs per environment that start from the same untrained student and differ only in the training seed. The three R2S runs share the one

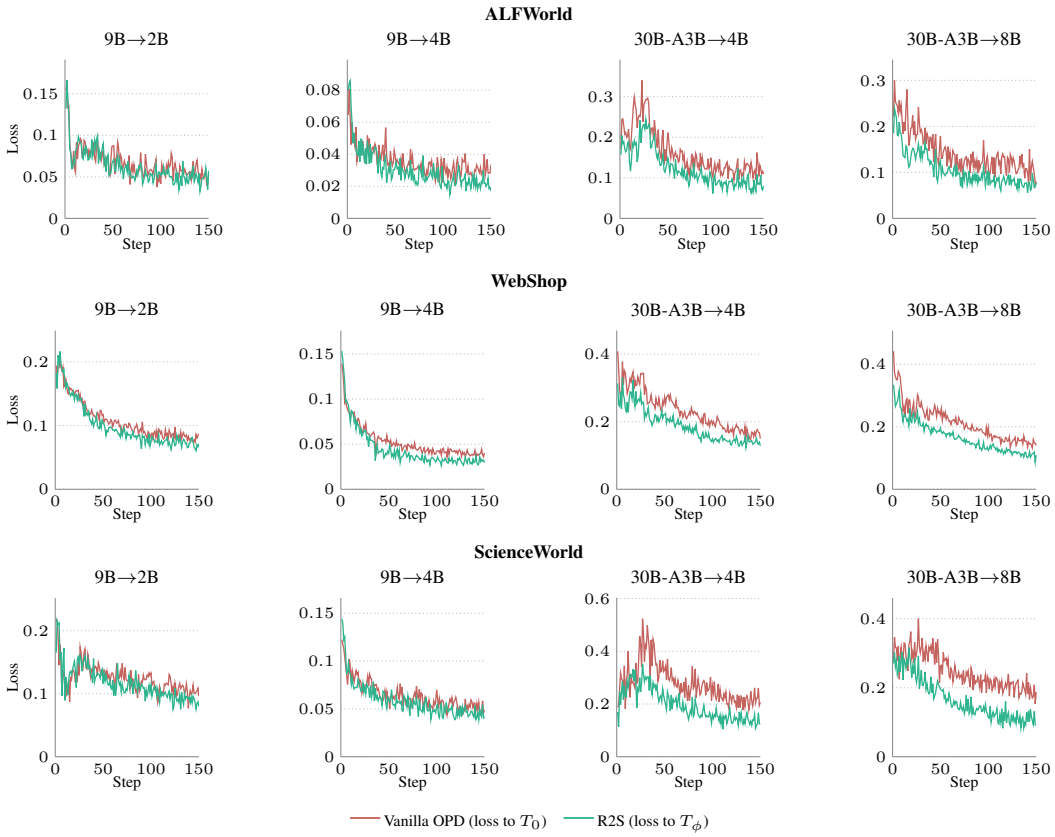

Figure 7: **Training loss of the same runs.** The sampled reverse KL of Eq. (4) to each run's own scorer (T_0 for vanilla OPD, T_ϕ for R2S), per optimizer update; the two scorers' losses are not comparable in level.

repaired teacher of Table 2, so its band reflects student training alone. In-training evaluation uses each run's own seed, so the bands also include evaluation sampling.

Baseline runs. TCOD-F2B, Guided-OPD and FutureBridge use the same models, prompt, optimizer settings, evaluation schedule and training seeds as the vanilla runs. Each of these four methods logs its own objective, plotted unsmoothed in Figure 8; vanilla OPD's sampled surrogate is not a nonnegative KL estimate, and Guided-OPD's mixes a teacher-token cross-entropy into the OPD term and FutureBridge's is computed on bridge segments over its curriculum, so their magnitudes are not comparable to the vanilla loss. Evaluation is the same for every method: full-horizon episodes on the same splits with the same budgets.

Curves and uncertainty. The horizontal axis counts completed optimizer updates. Each curve evaluates a method's trained student at updates 25, 50, 75, 100, 125, and 150. Update 0 is the same untrained student for every method, so all five curves start from the vanilla runs' evaluation of it. At each recorded update, the line is the arithmetic mean over the three runs and the shaded band is one sample standard deviation, using denominator $3 - 1$. No interpolation, missing-step imputation, or peak-checkpoint selection is applied.

Teacher-alone baseline. The dashed line is the mean success of the teacher acting alone under the same prompt, evaluation splits and averaging, temperature, and evaluation budgets; it is the anchor of Table 2.

Repeated evaluations. Four of the 252 in-training evaluations of vanilla OPD and the three baselines were repeated under the rule of Appendix D; each repeat replaces the original, and both records are kept.

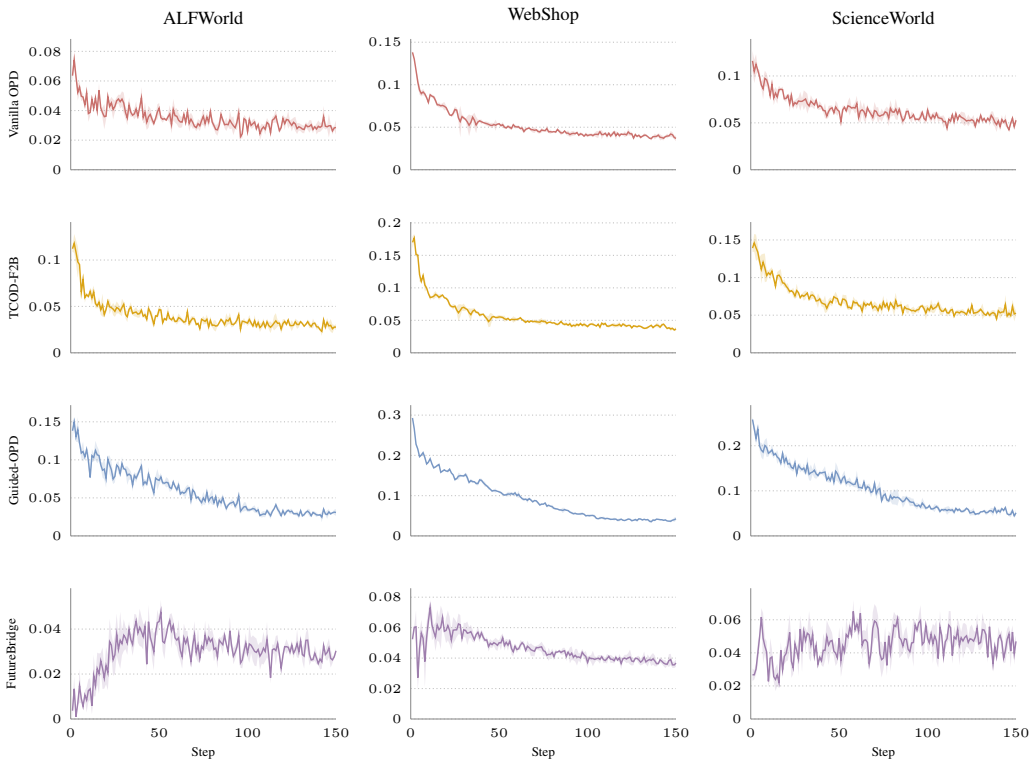

Figure 8: **Logged training objectives of vanilla OPD and the three baselines of Figure 3.** Mean $\pm$ one standard deviation over the same three training seeds, one row per method, one y scale per panel; magnitudes are not comparable across rows.

Table 18: **Validation target fidelity on the complete assistant suffix.** Each teacher's cross-entropy and top-16 KL against the same q^+ targets, on every assistant token of the validation states from the mined turn onward, averaged per state; lower is closer to the target.

	ALFWorld		WebShop		ScienceWorld	
Teacher	CE	KL	CE	KL	CE	KL
Original T_0	0.4660	0.0291	0.4272	0.0631	0.6516	0.0766
R2S	0.4493	0.0160	0.4004	0.0365	0.6122	0.0415
Label-SFT	0.4718	0.0349	0.4313	0.0671	0.6580	0.0837

Table 19: **Paired suffix target-fidelity differences: R2S minus each baseline teacher.** CE and KL as in Table 18; brackets: paired 95% goal-cluster bootstrap intervals.

Environment	Baseline	Δ CE [95% CI]	Δ KL [95% CI]
ALFWorld	Original T_0	-0.0167 [-0.0185, -0.0151]	-0.0131 [-0.0148, -0.0117]
ALFWorld	Label-SFT	-0.0225 [-0.0243, -0.0210]	-0.0190 [-0.0206, -0.0175]
WebShop	Original T_0	-0.0268 [-0.0315, -0.0222]	-0.0266 [-0.0313, -0.0221]
WebShop	Label-SFT	-0.0309 [-0.0357, -0.0263]	-0.0305 [-0.0353, -0.0260]
ScienceWorld	Original T_0	-0.0393 [-0.0444, -0.0346]	-0.0351 [-0.0399, -0.0304]
ScienceWorld	Label-SFT	-0.0458 [-0.0511, -0.0407]	-0.0422 [-0.0473, -0.0372]

G Label-SFT Control and Long-Horizon Evaluation

G.1 Label-SFT Control and Scorer Fidelity

Label-SFT control. Label-SFT fine-tunes the teacher on the same selected takeovers by standard SFT, for one epoch at the learning rate of the repair, and the resulting teacher is frozen for otherwise standard OPD.

Scorer fidelity. The benchmark of Tables 18 and 19 and of the left panel of Figure 4 holds the original student text and the cached q^+ distributions fixed across the three teachers, on every assistant position of the validation states from the mined turn onward; earlier context is retained without contributing a target. We report the cross-entropy of the full-vocabulary scorer probabilities against the top-16 renormalized targets, and the KL after renormalizing the scorer on the same support, averaged over tokens within a state and then over states, as in the repair objective. Table 19 gives 95% percentile intervals from 2,000 goal-cluster bootstrap resamples, paired across teachers; the whiskers in the left panel of Figure 4 are each teacher's own interval from the same resamples. Validation goals are withheld from the fine-tuning gradients.

A teacher fine-tuned to act better is not a better scorer. Table 20 follows the label-SFT teachers of ALFWorld and WebShop as actors and as scorers, for the 9B teacher's 4B students and for its 2B students, where a four-epoch dose was also run. In Table 20, *Held-out NLL* is the fine-tune's negative log-likelihood of held-out takeovers, before → after; *KL to* q^+ */* q^0 is the scorer's divergence, on the cell's own held-out states, from the informed pass at the student's thinking tokens and from the plain pass at its action tokens; *Acts alone* is the teacher acting under the evaluation prompt; *Student* is a student trained by vanilla OPD against that teacher under R2S's own protocol, read at the final checkpoint; success is the mean of the ALFWorld splits and the rate on the 500 canonical WebShop goals. The label-SFT teacher is a strong actor: every control teacher acts at least twelve points above the original teacher, and on ALFWorld at 2B both doses act above the repaired one. It is a worse scorer at every dose. Its likelihood of held-out takeovers falls, and falls further at the higher dose; its divergence from the informed pass exceeds the original teacher's and grows with the dose; and at the higher dose it moves the scores of the student's action tokens, which the repaired teacher leaves almost unchanged.

Its students improve on vanilla OPD but trail R2S's on WebShop. The last column of Table 20 trains a student by vanilla OPD against each control teacher, under R2S's own protocol, and reads the same three decodes of the final checkpoint; the 2B students take both doses, the 4B students,

Table 20: **A better actor is not a better scorer.** Label-SFT teachers fit the same selected takeovers as R2S by standard SFT. Held-out NLL before→after; KL to q^+ (thinking) and q^0 (action) on held-out states; success (%) of the teacher acting alone and of its student, mean±std over three decodes.

Teacher	Held-out NLL	KL to q^+	KL to q^0	Acts alone	Student	Student's loss
ALFWorld, 9B → 2B						
T_0 (original)	–	0.088	0.000	63.1±1.7	58.0±2.4	vanilla OPD
T_ϕ (R2S repair)	–	0.049	0.001	75.6±1.6	77.9±1.3	R2S
Label-SFT, 1 epoch, lr $2{\cdot}10^{-6}$	0.288 → 0.291	0.099	0.002	78.5±1.9	77.7±2.6	vanilla OPD
Label-SFT, 4 epochs, lr $2{\cdot}10^{-6}$	0.288 → 0.339	0.257	0.006	82.9±0.7	77.6±2.2	vanilla OPD
WebShop, 9B → 2B						
T_0 (original)	–	0.099	0.000	33.5±0.3	35.7±1.6	vanilla OPD
T_ϕ (R2S repair)	–	0.053	0.003	53.1±0.3	57.9±1.6	R2S
Label-SFT, 1 epoch, lr $2{\cdot}10^{-6}$	0.309 → 0.316	0.101	0.006	46.4±1.2	48.0±0.8	vanilla OPD
Label-SFT, 4 epochs, lr $2{\cdot}10^{-6}$	0.309 → 0.373	0.195	0.028	50.6±1.7	49.7±1.6	vanilla OPD
ALFWorld, 9B → 4B						
T_0 (original)	–	0.088	0.000	63.1±1.7	65.8±2.6	vanilla OPD
T_ϕ (R2S repair)	–	0.048	0.000	84.0±1.3	81.9±2.4	R2S
Label-SFT, 1 epoch, lr $2{\cdot}10^{-6}$	0.299 → 0.303	0.101	0.001	79.7±0.4	76.8±3.5	vanilla OPD
WebShop, 9B → 4B						
T_0 (original)	–	0.123	0.000	33.5±0.3	34.6±2.9	vanilla OPD
T_ϕ (R2S repair)	–	0.074	0.004	48.5±1.9	55.9±1.4	R2S
Label-SFT, 1 epoch, lr $2{\cdot}10^{-6}$	0.280 → 0.288	0.128	0.006	45.7±0.2	46.7±1.3	vanilla OPD

the label-SFT row of Table 5, the one-epoch dose. Every control student ends far above the vanilla student of its cell. On ALFWorld the control students match the R2S student at 2B and fall five points short of it at 4B, at or below their own teachers' acting level; on WebShop they land at their teachers' acting level, eight to ten points below the R2S student. A better actor does not make a better student here: at 2B on ALFWorld the four-epoch teacher acts seven points above the repaired teacher, and its student only matches the R2S student. The dose moves the control teacher's acting level and its divergence from the informed pass far more than it moves its students. Both repair and label-SFT benefit from fine-tuning on takeovers mined at the student's own divergence states.

G.2 Long-Horizon Behavior

Protocol. We compare Qwen3.5-4B students distilled from the original 9B teacher, the R2S teacher and the label-SFT teacher on ALFWorld, WebShop and ScienceWorld. Each student is the final checkpoint of its 150-update training run, decoded three times. The R2S students are the 9B-to-4B students of Table 2, re-decoded in a separate batch; the vanilla and label-SFT students reuse three decodes from an earlier batch of the same evaluation harness. Prompts, environments and each environment's turn budget H follow Appendix D, and each response generates at most 4,096 tokens at temperature 0.4. Figure 5 shows the environment-level curves in the main text; Tables 21 and 22 give every split and Table 23 the paired contrasts per split.

Environment-level contrasts. The paired R2S-minus-vanilla success differences are 16.7 points $[12.8, 20.7]$ on ALFWorld, 20.6 $[17.4, 23.8]$ on WebShop and 12.8 $[8.7, 17.0]$ on ScienceWorld, and the cumulative observed-failure-turn differences are -1.59 $[-1.93, -1.28]$, -0.40 $[-0.48, -0.32]$ and -0.68 $[-0.92, -0.44]$. Against label-SFT, success improves by 2.3 $[-0.9, 5.4]$, 9.5 $[6.9, 12.2]$ and 7.6 $[3.6, 11.5]$ points, and the failure-turn differences are -0.12 $[-0.29, 0.05]$, -0.05 $[-0.13, 0.03]$ and -0.19 $[-0.41, 0.04]$. Brackets are paired 95% goal-bootstrap intervals.

Cumulative observed errors and task outcomes. For goal g, decoding repetition r and turn u, let e_{gru} indicate a recorded invalid action or action-format error, counted once when both occur, following the observable interaction-failure diagnostic of SAGE-OPD (Zhou et al., 2026). With terminal turn L_{gr} and success flag z_{gr}, define

$$C_{gr}(t) = \sum_{u=1}^{\min(t, L_{gr})} e_{gru}, \qquad S_{gr}(t) = \mathbf{1}\{L_{gr} \le t\}\, z_{gr}, \qquad 0 \le t \le H. \tag{5}$$

Every initial task stays in the cohort after it ends: its error count stops growing, and its success or failure stays recorded. Success is the environment's completion flag, a full-match purchase on WebShop; the native task score of WebShop and ScienceWorld is reported separately.

Table 21: **Every split: task outcomes.** Means and 95% goal-bootstrap intervals; three decoding replicates averaged within goal. Success/timeout are percentages. Task score is the native normalized score, distinct from success and summed rewards. It is unavailable on ALFWorld.

Split	Method	Success (%)	Task score	Timeout (%)	Turns
ALF seen	Vanilla OPD	62.9 [56.4, 69.0]	–	37.1 [31.0, 43.6]	19.12 [17.75, 20.54]
	Label-SFT	76.2 [70.5, 81.4]	–	23.8 [18.6, 29.5]	16.14 [14.81, 17.50]
	R2S	79.5 [74.3, 84.8]	–	20.5 [15.2, 25.7]	14.95 [13.66, 16.25]
ALF unseen	Vanilla OPD	65.9 [59.2, 72.4]	–	34.1 [27.6, 40.8]	19.31 [17.87, 20.68]
	Label-SFT	81.3 [75.4, 86.3]	–	18.7 [13.7, 24.6]	16.06 [14.67, 17.43]
	R2S	82.6 [77.1, 87.6]	–	17.4 [12.4, 22.9]	15.45 [14.05, 16.76]
WS canonical	Vanilla OPD	36.3 [32.6, 39.7]	0.419 [0.381, 0.455]	54.4 [50.7, 58.5]	11.95 [11.64, 12.29]
	Label-SFT	47.3 [43.5, 51.1]	0.551 [0.512, 0.587]	39.3 [35.7, 43.3]	10.29 [9.96, 10.65]
	R2S	56.9 [53.0, 60.7]	0.659 [0.624, 0.692]	26.5 [23.3, 29.8]	8.99 [8.66, 9.34]
WS pool	Vanilla OPD	29.8 [25.0, 34.4]	0.366 [0.318, 0.415]	59.6 [54.6, 64.7]	12.26 [11.83, 12.68]
	Label-SFT	41.0 [35.8, 46.1]	0.505 [0.456, 0.558]	43.2 [38.0, 48.2]	10.53 [10.03, 11.01]
	R2S	54.0 [48.4, 59.4]	0.647 [0.600, 0.696]	26.7 [22.4, 31.1]	8.98 [8.53, 9.46]
SW seen	Vanilla OPD	37.1 [31.7, 42.9]	0.599 [0.556, 0.642]	33.8 [27.4, 40.0]	19.13 [17.73, 20.41]
	Label-SFT	48.8 [42.6, 55.0]	0.668 [0.621, 0.716]	22.6 [16.9, 28.3]	14.82 [13.25, 16.33]
	R2S	55.0 [48.8, 61.2]	0.734 [0.688, 0.778]	19.5 [14.0, 25.0]	15.25 [13.79, 16.59]
SW unseen	Vanilla OPD	31.3 [25.6, 37.6]	0.564 [0.513, 0.615]	30.8 [25.4, 36.3]	17.47 [16.05, 18.85]
	Label-SFT	30.1 [23.9, 36.1]	0.571 [0.516, 0.623]	33.8 [27.9, 39.6]	17.02 [15.40, 18.51]
	R2S	39.1 [33.3, 45.3]	0.639 [0.594, 0.685]	19.9 [15.7, 23.9]	13.79 [12.39, 15.13]

Reduction and uncertainty. We average the repetitions within each goal, then the goals within each split; ALFWorld and ScienceWorld average their two split means, and WebShop uses its 500 canonical goals. Paired contrasts subtract the baseline from R2S on the same goals. Intervals are 95% percentile intervals from 2,000 goal-bootstrap draws that keep each goal's repetitions together, resampled within each split before the splits are averaged.

Coverage. The cohort holds 11,736 episodes, 3,912 of the R2S students and 7,824 of the baseline students; every trajectory is retained.

Table 22: **Every split: cumulative observable errors.** Means and 95% goal-bootstrap intervals at horizons 30/15/30, including early exits. Failure turns count the union of invalid/format flags. First-invalid incidence is a percentage.

Split	Method	Invalid count	Format count	Fail turns	First invalid (%)
ALF seen	Vanilla OPD	1.42 [1.20, 1.65]	1.00 [0.69, 1.35]	2.42 [1.98, 2.88]	56.7 [50.2, 62.9]
	Label-SFT	0.84 [0.68, 1.01]	0.27 [0.13, 0.45]	1.11 [0.89, 1.36]	41.9 [36.0, 47.9]
	R2S	0.68 [0.53, 0.84]	0.09 [0.03, 0.17]	0.77 [0.60, 0.95]	37.1 [31.2, 43.3]
ALF unseen	Vanilla OPD	1.68 [1.43, 1.96]	0.96 [0.64, 1.32]	2.64 [2.19, 3.18]	60.9 [54.5, 67.4]
	Label-SFT	0.90 [0.74, 1.07]	0.09 [0.03, 0.19]	1.00 [0.82, 1.19]	45.5 [38.8, 52.2]
	R2S	0.98 [0.77, 1.20]	0.12 [0.04, 0.25]	1.10 [0.86, 1.37]	37.6 [31.6, 43.8]
WS canonical	Vanilla OPD	0.92 [0.86, 0.99]	0.00 [0.00, 0.00]	0.92 [0.86, 0.99]	59.7 [56.7, 62.9]
	Label-SFT	0.57 [0.52, 0.63]	0.00 [0.00, 0.00]	0.57 [0.52, 0.63]	42.2 [39.1, 45.2]
	R2S	0.52 [0.44, 0.61]	0.00 [0.00, 0.00]	0.52 [0.44, 0.61]	31.3 [28.3, 34.3]
WS pool	Vanilla OPD	0.90 [0.83, 0.98]	0.00 [0.00, 0.00]	0.90 [0.83, 0.98]	60.0 [56.0, 64.2]
	Label-SFT	0.58 [0.52, 0.65]	0.00 [0.00, 0.00]	0.58 [0.52, 0.65]	41.5 [37.5, 45.4]
	R2S	0.47 [0.40, 0.55]	0.00 [0.00, 0.00]	0.47 [0.40, 0.56]	32.0 [28.3, 36.1]
SW seen	Vanilla OPD	2.37 [1.97, 2.79]	0.10 [0.03, 0.20]	2.47 [2.07, 2.89]	57.6 [51.9, 63.6]
	Label-SFT	1.72 [1.40, 2.06]	0.05 [0.01, 0.10]	1.77 [1.43, 2.12]	51.0 [44.3, 57.4]
	R2S	1.52 [1.17, 1.89]	0.08 [0.01, 0.18]	1.60 [1.25, 1.96]	47.9 [41.2, 54.3]
SW unseen	Vanilla OPD	1.65 [1.35, 1.94]	0.26 [0.14, 0.41]	1.91 [1.58, 2.24]	50.7 [44.8, 56.5]
	Label-SFT	1.47 [1.19, 1.74]	0.16 [0.05, 0.30]	1.62 [1.30, 1.96]	44.0 [37.6, 50.0]
	R2S	1.23 [0.96, 1.52]	0.17 [0.05, 0.33]	1.41 [1.10, 1.71]	39.8 [33.8, 46.0]

Table 23: **Paired fixed-horizon contrasts: R2S minus each baseline.** Mean differences and 95% paired goal-bootstrap intervals. Positive success differences and negative failure-count differences favor R2S. All initial goals remain eligible.

Split	Baseline	Goals	Δ success (pp)	Δ fail turns
ALF seen	Vanilla OPD	140	+16.7 [+11.9, +21.9]	-1.65 [-2.13, -1.23]
	Label-SFT	140	+3.3 [-1.2, +8.1]	-0.34 [-0.59, -0.14]
ALF unseen	Vanilla OPD	134	+16.7 [+10.7, +22.9]	-1.54 [-2.06, -1.11]
	Label-SFT	134	+1.2 [-3.5, +6.0]	+0.10 [-0.13, +0.34]
WS canonical	Vanilla OPD	500	+20.6 [+17.3, +23.8]	-0.40 [-0.48, -0.32]
	Label-SFT	500	+9.5 [+6.9, +12.2]	-0.05 [-0.13, +0.03]
WS pool	Vanilla OPD	256	+24.2 [+20.4, +28.3]	-0.43 [-0.51, -0.36]
	Label-SFT	256	+13.0 [+9.6, +16.5]	-0.11 [-0.19, -0.02]
SW seen	Vanilla OPD	140	+17.9 [+11.7, +23.8]	-0.87 [-1.20, -0.55]
	Label-SFT	140	+6.2 [+0.2, +12.1]	-0.17 [-0.42, +0.08]
SW unseen	Vanilla OPD	134	+7.7 [+1.2, +14.2]	-0.50 [-0.84, -0.13]
	Label-SFT	134	+9.0 [+3.2, +14.7]	-0.22 [-0.57, +0.16]