TAAD: TEACHABILITY-AWARE AGENTIC DISTILLATION FOR MULTI-STEP CREDIT ASSIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-step agentic reinforcement learning trains agents to complete long-horizon trajectories under a sparse terminal return, where Group Relative Policy Optimization (GRPO) broadcasts one sequence-level advantage uniformly to every token. Token-level teacher-student self-distillation sharpens this signal, yet single-token probability gaps can conflate distributional disagreement with incidental fluctuation and do not exploit step structure. In this paper, we propose Teachability-Aware Agentic Distillation (TAAD), a bounded two-level advantage reweighting method. TAAD replaces the single-token gap with a teacher-mass-weighted disagreement signal over the shared teacher-student top-K support, aggregates this signal over action spans, and gates its rising inter-step transitions by student action surprisal to form a step-priority score. A maximum-entropy softmax step budget is then refined into per-token weights. The reshaped advantage preserves the sign of the sequence-level advantage at every valid token while allocating different credit magnitudes across a trajectory. Across both Qwen2.5-3B-Instruct and Qwen2.5-7B-Instruct backbones, TAAD consistently outperforms all baselines on the three multi-step benchmarks of ALFWorld, Search-QA, and WebShop. Code is here.

1 INTRODUCTION

Multi-step agentic reinforcement learning trains a policy to interact with an environment through sequences of reasoning, tool use, and actions while optimizing task outcomes (Wang et al., 2025; Chen et al., 2025). This setting underlies embodied control (Shridhar et al., 2020), web interaction (Yao et al., 2022), and search-based question answering (Jin et al., 2025), where an early local decision can constrain every subsequent observation and action. Group Relative Policy Optimization (GRPO) provides a critic-free recipe for such tasks by converting group outcomes into trajectory-level advantages (Shao et al., 2024). Yet its central optimization problem is inherently structural. A sparse terminal return must assign credit across heterogeneous decisions, though only a small subset may determine whether the trajectory succeeds (Feng et al., 2025; Zhang et al., 2026).

Recent work has sought denser supervision by combining outcome-grounded reinforcement learning with on-policy self-distillation. RLSD retains the environment reward as the update-direction signal and uses teacher-student discrepancies to modulate update magnitudes (Yang et al., 2026a). SDAR instead introduces a gated auxiliary distillation objective to exploit privileged teacher guidance during multi-turn agent training (Lu et al., 2026), and step-aware variants further segment trajectories into action-centered spans to shape advantages at the step level (Zhang et al., 2026). These methods establish that teacher-derived signals can refine uniform trajectory credit without replacing the reinforcement learning backbone. However, their credit signals remain primarily token-local, use a static per-step allocation, or rely on log-probability gaps. These choices do not explicitly distinguish a persistent disagreement level from where disagreement begins to rise across a trajectory.

This distinction matters in long-horizon interaction. Prior work reports that errors compound across steps and make teacher guidance increasingly unreliable (Wang et al., 2026b; Zhong et al., 2026). A high discrepancy at one step need not identify a uniquely informative intervention point, because it may simply persist from an earlier decision. Conversely, a sharp rise in disagreement marks a transition where the action distributions diverge more than at the preceding step under privileged teacher context. Single-token gaps are also sensitive to incidental probability fluctuations (Fu et al., 2026), while a teacher preference outside the student support is excluded by a shared-support com-

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.

parison. The challenge has two coupled parts. First, the method needs a conservative token-level signal that emphasizes disagreement on candidates both policies consider. Second, it must convert this local signal into a trajectory-aware allocation that prioritizes rising transitions without declaring non-rising steps uninformative. Uniform or static credit rules do not encode this transition-level distinction, and unconstrained reweighting can obscure the reliable sign from environment feedback.

In this paper, we introduce **Teachability-Aware Agentic Distillation (TAAD)**, a bounded two-level advantage reweighting framework for multi-step agentic RL. TAAD first constructs a teacher-mass-weighted disagreement signal over the shared teacher-student top-K support. This construction compares distributions rather than a sampled-token gap and suppresses disagreement lacking candidate support. TAAD then aggregates it within causal action spans and detects rising transitions across steps. Student action surprisal gates this transition score, yielding a step-priority score over action spans with high sampled surprisal and increasing shared-support disagreement. A maximum-entropy softmax converts the priorities into a positive step budget, and TAAD refines the budget within each step using normalized token-level disagreement. Clipped token weights then reshape the GRPO advantage while preserving the sequence-level advantage sign at each valid token, leaving the environment reward to determine whether sampled behavior is reinforced or penalized. These mechanisms recast teacher guidance as bounded transition-priority allocation rather than direct policy imitation or static token selection. Our contributions are as follows:

- We formulate multi-step agent credit assignment as a transition-priority allocation problem and introduce TAAD, which converts a sparse sequence-level GRPO advantage into bounded step-token credit while retaining environment-derived update signs.
- We develop a two-level mechanism that combines teacher-mass-weighted shared-support disagreement, action-surprisal-gated rising transitions, and maximum-entropy step budgeting to distinguish onset-aware priority from static disagreement levels.
- We evaluate TAAD on ALFWorld, Search-QA, and WebShop with both Qwen2.5-3B-Instruct and Qwen2.5-7B-Instruct, consistently outperforming all baselines on aggregate metrics.

2 Related Work

Agentic Reinforcement Learning and Credit Assignment. Agentic reinforcement learning trains language models to perform multi-turn reasoning, search, and tool use through feedback from the environment. RLVR methods, exemplified by GRPO, DeepSeek-R1, and DAPO, optimize reasoning policies using verifiable terminal rewards (Shao et al., 2024; DeepSeek-AI, 2025; Yu et al., 2025) and have been further extended to search, multi-turn decision-making, hierarchical interaction, and general-purpose agent training (Jin et al., 2025; Wang et al., 2025; Chen et al., 2025; Zhou et al., 2024; Feng et al., 2025; Luo et al., 2025). However, these methods typically construct a trajectory-level advantage from the final outcome and broadcast the same advantage to all tokens in the response, making it difficult to distinguish critical actions from ineffective reasoning and formatting-related content. Process supervision, implicit process rewards, and outcome-grounded advantage reshaping seek to provide more fine-grained learning signals for intermediate decisions (Lightman et al., 2024; Cui et al., 2025; Li et al., 2026), but they often rely on additional evaluators, process-level annotations, or heuristic credit proxies. Consequently, action-level credit assignment in long-horizon interactions remains challenging.

On-Policy Self-Distillation with Privileged Information. On-Policy Distillation (OPD) queries the teacher on student-generated rollouts, reducing the distribution mismatch of offline distillation (Agarwal et al., 2024; Gu et al., 2024). OPSD and related methods use privileged information, such as reference answers, environmental feedback, and successful trajectories, to construct self-teachers that provide dense token-level supervision (Zhao et al., 2026; Hübotter et al., 2026; Penaloza et al., 2026; Shenfeld et al., 2026; He et al., 2026; Wang et al., 2026a; Ye et al., 2026; 2025; Yang et al., 2026b). To improve signal imbalance, prefix drift, and instability in sampled-token OPD, prior work has explored top-K local-support matching, entropy-aware KL objectives, important-token selection, and adaptive target reformulation (Fu et al., 2026; Jin et al., 2026; Xu et al., 2026; Jang et al., 2026). RLSD uses environmental rewards to determine update directions and teacher signals to modulate their magnitudes (Yang et al., 2026a), whereas SDAR introduces gated self-distillation as an auxiliary objective for multi-turn agent RL (Lu et al., 2026).

Step-Level and Long-Horizon Credit Allocation. To better align learning signals with agent interaction structures, recent studies have moved from trajectory- or token-level optimization toward turn-, step-, and action-level modeling. Hierarchical multi-turn RL, local state grouping, and process-reward methods estimate local contributions using sub-trajectories, shared states, or intermediate outcomes (Chen et al., 2025; Zhou et al., 2024; Feng et al., 2025; Lightman et al., 2024; Cui et al., 2025; Li et al., 2026). SkillRL supports long-horizon learning through recursive skill augmentation (Xia et al., 2026), TCOD schedules distillation across interaction stages (Wang et al., 2026b), SOD dynamically reweights supervision according to step-level teacher-student divergence (Zhong et al., 2026), and StepOPSD segments trajectories into action-centered spans to shape step-level advantages (Zhang et al., 2026). Although these methods reduce credit ambiguity and error accumulation, they often rely on fixed curricula, auxiliary signals, or uniform weights within each step, leaving cross-step importance and within-step token contributions insufficiently modeled.

3 PRELIMINARIES

We consider multi-step agentic reinforcement learning where an agent interacts with an environment over a trajectory and receives a sparse return R at trajectory termination; TAAD does not introduce any additional process reward. Let π_θ denote the student policy. Following GRPO (Shao et al., 2024), for a prompt x we sample a group of G trajectories $\{\tau_i\}_{i=1}^{G}$ from the old policy $\pi_{\theta_{\text{old}}}$, and the sequence-level advantage of trajectory τ_i is the group-normalized reward

$$A_{\text{seq}}^{(i)} = \frac{R(\tau_i) - \mu_R}{\sigma_R + \epsilon}, \tag{1}$$

where μ_R and σ_R are the mean and standard deviation of the group rewards. GRPO broadcasts this single scalar uniformly to every token, which we aim to refine. In what follows we write A_{seq} for the advantage of the trajectory under consideration, dropping the superscript (i).

Following prior self-distillation work (Lu et al., 2026; Yang et al., 2026a), we adopt a teacher-student scheme with the same architecture. The teacher parameters $\bar{\theta}$ are periodically synchronized from the student and frozen between synchronizations, and the teacher conditions on privileged information r (a retrieved skill) that is available only during training. We write

$$\pi_S(v) = \pi_{\theta_{\text{old}}}(v \mid x, y_{<t}), \qquad \pi_T(v) = \pi_{\bar{\theta}}(v \mid x, r, y_{<t}), \tag{2}$$

for the student and teacher next-token distributions at position t, evaluated at the rollout parameters θ_{old} and frozen snapshot $\bar{\theta}$, respectively. We detach and retain these statistics across the GRPO epochs. The environment reward supplies the sequence-level advantage sign. These fixed statistics modulate its per-token magnitude. Each action span consists of the response tokens between two consecutive environment observations, so steps align with environment turns; empty spans are discarded. We index the spans of a trajectory by $s \in \{1, \ldots, N\}$, and each step contains at least one valid response token. Let $\text{mask}_t \in \{0, 1\}$ indicate a valid response token. Both policies retain their top-K candidate tokens and probabilities; concrete settings are deferred to the experiments.

4 METHODOLOGY

4.1 OVERVIEW

As illustrated in Figure 1, TAAD reframes credit assignment as a transition-priority allocation problem and comprises three stages. Stage 1 measures M_t, a teacher-mass-weighted disagreement signal over shared top-K support, ensuring the token-level signal reflects distributional disagreement rather than incidental fluctuation. Crucially, Stage 2 locates the rising transitions: it forms a transition-priority score q_s from rising changes of the step-aggregated signal gated by normalized action surprisal, and allocates the step budget b_s by softmax. Stage 3 refines the budget within each step into a per-token weight $w_{s,t}$, which is clipped and blended with the sequence-level advantage, ultimately reallocating credit magnitudes while preserving the outcome-derived update direction.

4.2 ROBUST TOKEN-LEVEL SHARED-SUPPORT DISAGREEMENT

A single-token log-probability gap $\Delta_t = \log \pi_T(y_t) - \log \pi_S(y_t)$ can be sensitive to distributional fluctuation (Yang et al., 2026a). We instead compare the teacher and student top-K distributions.

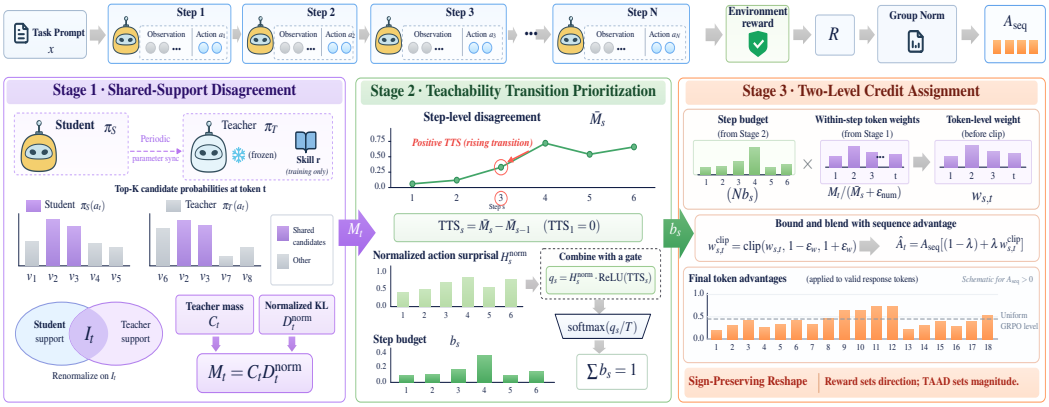


Figure 1: Overview of TAAD. The environment reward yields the sequence-level advantage A_{seq} whose sign is preserved; teacher guidance only reshapes per-token credit magnitudes. Stage 1 computes the teacher-mass-weighted disagreement M_t on the shared top-K support. Stage 2 aggregates M_t into $\overline{M}_s$, detects rising transitions via surprisal-gated TTS_s, and allocates a maximum-entropy step budget b_s. Stage 3 combines the budget with within-step token weights, yielding bounded per-token advantages $\hat{A}_t$ with reallocated magnitudes and an unchanged sign.

Intersection and Shared-Support Mass. Let $I_t = \text{TopK}_S(t) \cap \text{TopK}_T(t)$ be the tokens ranked in the top-K of both policies. We measure the teacher probability mass on this shared support as

$$C_t = \sum_{v \in I_t} \pi_T(v \mid x, r, y_{<t}), \tag{3}$$

which is large when the shared candidates retain substantial teacher probability. A large C_t alone is not enough, since the teacher may still agree with the student after renormalization on that support.

Renormalized Divergence. To quantify how the teacher and student disagree on the same candidates, we renormalize both distributions over the intersection and compute the forward KL

$$\bar{p}_T^{I_t}(v) = \frac{\pi_T(v)}{\sum_{u \in I_t} \pi_T(u)}, \quad \bar{p}_S^{I_t}(v) = \frac{\pi_S(v)}{\sum_{u \in I_t} \pi_S(u)}, \quad D_t = D_{\text{KL}}\left(\bar{p}_T^{I_t} \,\|\, \bar{p}_S^{I_t}\right). \tag{4}$$

Restricting to the intersection filters out long-tail candidates that neither policy endorses, so D_t rises only when the two policies systematically disagree on shared candidates. If the intersection contains fewer than two tokens, we set $D_t = 0$, since the renormalized KL is degenerate.

Robust Normalization and Merging. We clip D_t by batch quantiles to suppress outliers,

$$D_t^{\text{norm}} = \text{clip}\left(\frac{D_t - Q_{5\%}(D)}{Q_{95\%}(D) - Q_{5\%}(D) + \epsilon},\ 0,\ 1\right), \tag{5}$$

where the quantiles are computed over all valid response tokens in the current batch and ϵ guards against degenerate batches with identical D_t. The token-level shared-support disagreement signal is

$$M_t = D_t^{\text{norm}} \cdot C_t \cdot \mathbf{1}[\text{mask}_t]. \tag{6}$$

A position has large M_t only when the teacher is confident on the shared candidates and the two policies disagree on their ranking and weighting, forming a conservative, teacher-mass-weighted disagreement signal. Together with the top-K student outputs already required by GRPO, all three ingredients I_t, C_t, and D_t are lightweight post-processing of one additional teacher forward pass.

4.3 Teachability Transition Signal

Moving from tokens to steps, we first aggregate M_t within each step into a step-level disagreement,

$$\overline{M}_s = \frac{\sum_{t \in \text{step}_s} M_t \cdot \text{mask}_t}{\sum_{t \in \text{step}_s} \text{mask}_t}, \tag{7}$$

and measure the student's surprise at its sampled actions via the within-step mean surprisal

$$H_s = -\frac{\sum_{t \in \text{step}_s} \log \pi_S(y_t) \cdot \text{mask}_t}{\sum_{t \in \text{step}_s} \text{mask}_t}. \tag{8}$$

Note that H_s is a conditional action surprisal over the tokens of step s; it is distinct from the step-budget entropy $H(b)$ used in Stage 2, which measures how the allocated budget is spread across steps. We normalize H_s with batch quantiles into $H_s^{\text{norm}} \in [0, 1]$, where the quantiles are computed over all valid steps in the current batch. Since the actions are sampled from the old policy, we use cached old-policy log-probabilities rather than recomputing them during the GRPO epochs.

Transition Signal. A static per-step score cannot locate the onset of rising disagreement. We therefore prioritize its change across steps, defining the Teachability Transition Signal (TTS). To retain a backward difference at the terminal step while avoiding a non-comparable score, we set

$$\text{TTS}_1 = 0, \qquad \text{TTS}_s = \overline{M}_s - \overline{M}_{s-1} \quad (s \geq 2). \tag{9}$$

A positive TTS_s marks an onset of rising shared-support disagreement. This is an onset-aware priority, not a claim that non-rising steps contain no teacher knowledge. The backward form also gives the terminal step a nonzero score, whereas a forward difference would force $\text{TTS}_N = 0$. Since $M_t \in [0, 1]$ and hence $\overline{M}_s \in [0, 1]$, the difference lies in $[-1, 1]$; this boundedness alone does not imply robustness to noisy step estimates. When steps are fine-grained, one may smooth $\overline{M}_s$ with a window of size three before differencing, an implementation option.

Transition-Aware Step Score. We gate the rising transition by student action surprisal,

$$q_s = H_s^{\text{norm}} \cdot \text{ReLU}(\text{TTS}_s). \tag{10}$$

A step obtains a high q_s only when the student is uncertain and shared-support disagreement is rising. A non-rising step receives no transition bonus, but is not thereby interpreted as lacking teacher knowledge. The outcome itself is handled by the sign of A_{seq} in the final advantage, so the transition detector does not branch on success or failure.

Step Budget. We allocate the step budget within a trajectory by softmax,

$$b_s = \frac{\exp(q_s/T)}{\sum_{s' \in \tau} \exp(q_{s'}/T)}, \tag{11}$$

with temperature T (default 1.0). This is the exact solution to a linear transition-priority utility with a maximum-entropy regularizer (Proposition 1), where T trades off concentrating budget on high-priority transitions against retaining allocation over the remaining steps. Softmax assigns every step positive budget. We report the trajectory-normalized entropy $H(b)/\log N$ over trajectories with $N > 1$. We exclude single-step trajectories because $\log N = 0$ and no cross-step allocation exists.

4.4 Two-Level Credit Assignment

Stage 2 delivers the step-level allocation: transitions scored into q_s and budgeted into b_s with $\sum_s b_s = 1$. We now refine this budget within each step. A direct normalization $w_{s,t} = b_s M_t / \sum_u M_u$ has magnitude $\mathcal{O}(1/(NL))$, where L is the number of response tokens. Renormalizing it over a trajectory would move all weights toward one, weakening the distinction between transition steps. We instead use a multiplicative form whose scale is compatible with clipping.

Within-Step and Across-Step Normalization. We normalize the token divergence within each step by its valid-token mean, so that the resulting weight is invariant to the step length,

$$\tilde{M}_{s,t} = \frac{M_t \cdot \text{mask}_t}{\frac{1}{\sum_{u \in \text{step}_s} \text{mask}_u} \sum_{u \in \text{step}_s} M_u \cdot \text{mask}_u + \epsilon_{\text{num}}}, \tag{12}$$

which is consistent with the $\overline{M}_s$ aggregation. We standardize the step budget to unit mean, $\tilde{b}_s = N \cdot b_s$, which has unit mean because $\sum_s b_s = 1$. Here N is the number of steps in the current

trajectory, so the standardization is per-trajectory, and the subsequent ε_w clip bounds any cross-trajectory scale difference. The per-token weight is their product

$$w_{s,t} = \tilde{b}_s \cdot \tilde{M}_{s,t}. \tag{13}$$

Ignoring the numerical stabilizer ϵ_{num}, both factors are normalized to unit mean over their respective step or trajectory domains. With $\epsilon_{\text{num}} > 0$, this is approximate; an all-zero-M_t step is consistently down-weighted before clipping. A transition step with a high b_s combined with a high $\tilde{M}_{s,t}$ receives $w_{s,t} > 1$, while an uninformative position receives $w_{s,t} < 1$.

Final Advantage. We clip and fold the weight into the sequence-level advantage,

$$w_{s,t}^{\text{clip}} = \text{clip}(w_{s,t},\ 1-\varepsilon_w,\ 1+\varepsilon_w), \qquad \hat{A}_t = A_{\text{seq}} \left[(1-\lambda) + \lambda\, w_{s,t}^{\text{clip}} \right], \tag{14}$$

where $\lambda \in [0,1]$ mixes the TAAD signal with uniform GRPO credit and ε_w bounds the per-token deviation. For each response token, let $\rho_t(\theta) = \pi_\theta(y_t \mid x, y_{<t}) / \pi_{\theta_{\text{old}}}(y_t \mid x, y_{<t})$ denote the policy likelihood ratio. We optimize the standard clipped surrogate

$$\mathcal{L}_{\text{TAAD}}(\theta) = -\mathbb{E}_t \Big[\min \Big(\rho_t(\theta)\hat{A}_t,\ \text{clip}(\rho_t(\theta), 1-\varepsilon_p, 1+\varepsilon_p)\hat{A}_t \Big) \Big] + \beta\, \mathcal{L}_{\text{KL}}, \tag{15}$$

where ε_p is the PPO clipping radius, β is the KL coefficient, $\mathcal{L}_{\text{KL}}$ is the GRPO KL penalty, and the expectation averages the surrogate over all valid response tokens in the mini-batch. The teacher parameters are synchronized with the student every K_{sync} optimizer steps and frozen in between. Setting $\lambda = 0$ recovers GRPO. In practice, we anneal λ linearly from λ_0 to zero over a fixed fraction of training, which smoothly hands the policy back to the outcome-grounded objective as the teacher signal becomes less informative; the initial value is examined in the sensitivity analysis.

For $\varepsilon_w < 1$, $w_{s,t}^{\text{clip}} \in [1-\varepsilon_w, 1+\varepsilon_w]$, so the reshaped advantage preserves the sign of A_{seq} at each valid token with deviation at most $\lambda\varepsilon_w|A_{\text{seq}}|$: TAAD is a bounded advantage reweighting. We collect the max-entropy derivation of the softmax budget (Proposition 1) and its separation from uniform and importance-based allocation (Proposition 2) in Appendix A, with proofs in Appendix B.

5 Experiments

5.1 Experimental Setup

Benchmarks. Following recent work on self-distillation (Yang et al., 2026a; Lu et al., 2026), we evaluate Qwen2.5-3B-Instruct and Qwen2.5-7B-Instruct on three multi-step environments. ALFWorld (Shridhar et al., 2020) is a text-based embodied environment with long-horizon tasks, and we report success rates across its six task families (Pick, Look, Clean, Heat, Cool, and Pick2). Search-QA follows Search-R1 (Jin et al., 2025): the agent retrieves external evidence before answering questions from NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), PopQA (Mallen et al., 2023), HotpotQA (Yang et al., 2018), 2Wiki (Ho et al., 2020), MuSiQue (Trivedi et al., 2022), and Bamboogle (Press et al., 2023). We report the answer accuracy for each dataset and the average over all instances. WebShop (Yao et al., 2022) is an interactive online-shopping environment with multi-page user requests, and we report the normalized task score and success rate.

Baselines. We compare against three categories. (1) *Training-free methods*: Vanilla evaluates the instruction-tuned model, and Skill-Prompt augments the input with retrieved skills at inference time. (2) *Outcome-based RL and general self-distillation*: GRPO (Shao et al., 2024) optimizes verifiable trajectory outcomes; OPSD (Zhao et al., 2026) and RLSD (Yang et al., 2026a) introduce self-distillation signals without constructing reusable skills. (3) *Skill-based self-distillation*: Skill-SD (Wang et al., 2026a) and SDAR (Lu et al., 2026) incorporate skill-conditioned privileged guidance into policy optimization. All methods use matched Qwen2.5-Instruct backbones and benchmark protocols, and skill-conditioned methods share the same SkillBank and retrieval protocol.

Implementation Details. All methods use VeRL with FSDP and vLLM on eight A100 GPUs under matched environment configurations. We use AdamW with learning rate 1×10^{-6}, GRPO group size $G = 8$, KL coefficient 0.01, and mini-batch size 256 turns. ALFWorld uses 16 tasks per

Table 1: Main results across two Qwen2.5-Instruct backbones. We report success rate (%) on ALFWorld, answer accuracy (%) on Search-QA, and normalized score and success rate (%) on WebShop. **Best** results are in bold, and second-best results are underlined.

Method	ALFWorld							Search-QA								WebShop	
	Pick	Look	Clean	Heat	Cool	Pick2	Avg.	NQ	Triv	Pop	Hotp	2Wk	MuS	Bam	Avg.	Score	Succ.
Backbone: Qwen2.5-3B-Instruct																	
Vanilla	18.8	37.5	8.3	6.2	4.2	8.3	11.7	16.9	33.8	28.0	9.5	14.3	2.5	30.4	21.3	11.5	4.7
Skill-Prompt	21.9	25.0	8.3	18.8	4.2	8.3	13.3	24.9	39.8	27.5	15.2	15.9	3.3	24.8	24.3	14.1	6.2
OPSD	15.6	37.5	16.7	6.2	16.7	16.7	16.4	1.4	0.9	1.1	1.9	1.2	1.7	0.8	1.2	13.7	5.5
GRPO	90.6	**75.0**	83.3	56.2	66.7	54.2	72.7	39.9	58.1	41.2	36.7	33.9	11.4	56.8	41.0	76.6	61.7
Skill-SD	**93.8**	62.5	79.2	**81.2**	58.3	54.2	73.4	42.4	58.9	42.6	38.1	35.2	12.2	**58.4**	42.3	73.4	61.7
RLSD	**93.8**	**75.0**	79.2	62.5	58.3	66.7	74.2	41.9	57.9	41.5	37.3	33.5	12.6	**58.4**	41.3	78.9	66.4
SDAR	**93.8**	62.5	**91.7**	75.0	**70.8**	66.7	79.7	45.6	**60.4**	**42.9**	42.9	36.6	14.8	28.0	44.0	80.5	63.3
TAAD (Ours)	**93.8**	**75.0**	87.5	**81.2**	**70.8**	**79.2**	**82.8**	**46.2**	58.6	42.3	**43.0**	**41.4**	**17.7**	26.4	**44.8**	**82.8**	**72.7**
Backbone: Qwen2.5-7B-Instruct																	
Vanilla	34.4	25.0	12.5	6.2	4.2	0.0	14.1	17.3	34.5	25.4	11.6	16.4	3.5	32.8	21.7	5.5	0.8
Skill-Prompt	37.5	37.5	16.7	12.5	8.3	4.2	18.8	22.4	36.4	31.0	16.7	14.7	4.5	30.4	24.3	10.9	3.9
OPSD	56.2	37.5	20.8	25.0	12.5	12.5	28.1	2.2	2.5	2.4	1.9	2.9	1.2	2.4	2.4	6.6	2.3
GRPO	93.8	**75.0**	75.0	62.5	70.8	66.7	75.8	46.0	61.0	41.9	39.4	37.0	15.6	58.4	43.6	78.1	68.0
Skill-SD	87.5	62.5	91.7	**81.2**	58.3	62.5	75.8	43.5	60.1	42.6	39.6	37.6	15.4	56.8	43.6	79.7	67.2
RLSD	93.8	**75.0**	79.2	68.8	66.7	58.3	75.0	**47.4**	61.2	41.9	40.8	34.8	15.8	58.4	43.4	81.2	70.3
SDAR	93.8	62.5	91.7	**81.2**	70.8	**75.0**	82.0	44.6	61.4	41.2	41.6	40.3	17.0	57.6	44.6	83.6	71.1
TAAD (Ours)	**100.0**	**75.0**	**95.8**	**81.2**	**75.0**	**75.0**	**85.9**	45.3	**62.4**	**47.0**	**42.6**	**42.7**	**18.7**	**60.0**	**47.3**	**85.9**	**78.1**

batch, 2,048/512 prompt/response tokens, and at most 50 steps, as in GIGPO (Feng et al., 2025). Search-QA uses E5 retrieval (Wang et al., 2022), 128 tasks per batch, 4,096/512 tokens, and at most four steps. WebShop uses a 1,000-task pool, 16 tasks per batch, 4,096/512 tokens, and at most 15 steps. Unless otherwise specified, TAAD uses $K = 10$, $T = 1.0$, $\lambda_0 = 0.5$, and clipping bounds $\varepsilon_w = \varepsilon_p = 0.2$; the frozen teacher is synchronized every $K_{\text{sync}} = 10$ optimizer steps.

5.2 Main Results

Table 1 presents the main results of all compared methods on the three benchmarks under both backbones, and we highlight three key observations below.

Overall Performance. TAAD achieves the strongest aggregate result on every benchmark with both backbones, consistently outperforming training-free prompting and outcome-only GRPO, indicating the improvement originates from the learning objective rather than inference-time skill augmentation. Its magnitude tracks the environment structure: the margin widens as the horizon lengthens and the decisions become more tightly interdependent, since longer trajectories offer more steps for bounded credit reallocation to separate informative transitions from routine ones.

Advances Over Teacher-Guided Baselines. TAAD outperforms general and skill-based self-distillation methods on all aggregate metrics. OPSD, optimizing distillation without outcome grounding, fails to improve over vanilla on Search-QA, corroborating the TAAD design in which rewards determine update signs. The consistent advantage over outcome-grounded teacher-guided methods aligns with the formulation: privileged teacher information enters as a bounded redistribution of the sequence-level advantage over rising transitions and shared-support disagreement, rather than an unconstrained auxiliary target or a static per-step weight. With matched backbones, protocols, and skill retrieval, the comparison isolates the credit assignment objective.

Heterogeneous Fine-Grained Results. The improvement is broad yet varies across fine-grained task families and datasets. On ALFWorld task families and Search-QA datasets, TAAD sometimes ties or trails the strongest baseline while still delivering the best overall averages, whereas its improvement on WebShop is consistent across both metrics. This heterogeneity echoes the transition-priority formulation: bounded step-token reweighting reallocates credit magnitudes within a trajectory and targets rising transitions rather than optimizing any particular task family, so the aggregate improvement stems from the joint two-level allocation accumulated across environments.

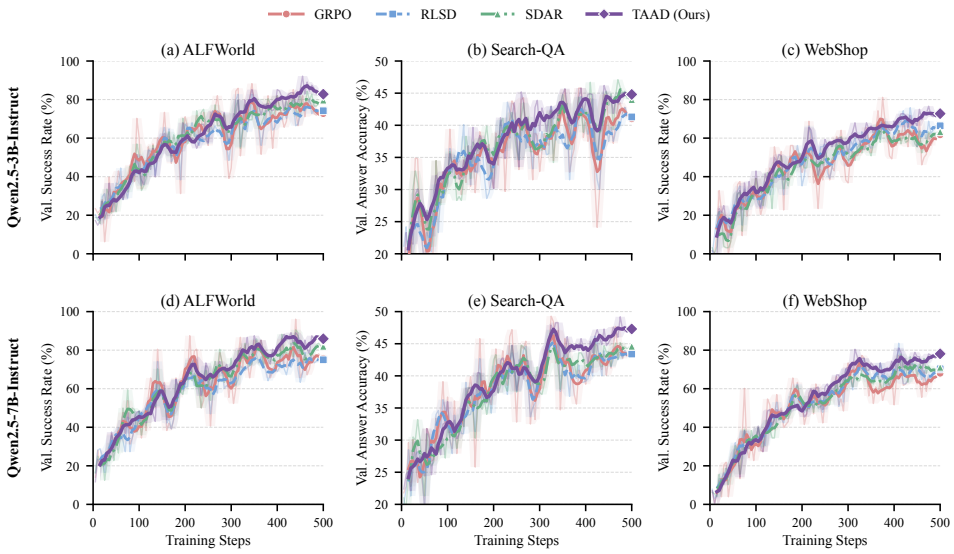


Figure 2: Validation curves for Qwen2.5-3B-Instruct (top) and Qwen2.5-7B-Instruct (bottom) on the three benchmarks. Faint traces show raw evaluations recorded every five training steps, solid traces show centered moving averages, and shaded envelopes show the local raw range.

Table 2: Ablation studies with Qwen2.5-3B-Instruct. Each variant removes one mechanism from the full method under the matched training and evaluation configuration.

Variant	ALFWorld Avg.	Search-QA Avg.	WebShop Succ.
TAAD (full)	**82.8**	**44.8**	**72.7**
Teachability Estimation			
w/o Shared-Support Signal	76.6	42.7	67.2
w/o Teacher Mass	80.5	43.6	69.5
Transition Prioritization			
w/o Transition Awareness	77.3	43.0	65.6
Hierarchical Credit Allocation			
w/o Step-Level Allocation	78.9	43.5	68.8
w/o Token-Level Refinement	81.2	44.6	70.3

5.3 Training Dynamics

Figure 2 complements the final results with the corresponding validation curves. TAAD attains the strongest result for both backbones ahead of competitive baselines. Notably, this advantage is built up rather than present initially: the TAAD curve starts alongside the leading baselines, steadily closes the early gap, and then pulls ahead as optimization proceeds, with the margin still widening at the end of training. This gradual takeoff is precisely what the bounded design intends: by preserving outcome-derived update directions and reallocating only credit magnitudes, TAAD keeps early training as stable as GRPO while letting targeted allocation compound into a growing lead, rather than trading early gains for later stagnation or instability.

5.4 Ablation Studies

Table 2 quantifies the contribution of each mechanism by removing one component at a time from the full method. Every ablation degrades all three aggregate metrics, indicating that the improvement stems from the joint effect of the three stages rather than from any single component. The two signal-oriented mechanisms are the most critical: removing the shared-support construction or the transition awareness causes the largest degradations, suggesting that both the quality of the teachability signal and the ability to locate its rising transitions are essential to the gains. Removing the teacher mass or either level of the hierarchical allocation yields smaller yet consistently negative effects, with the step-level budget contributing more than the within-step token refinement. This ordering echoes the design of TAAD: what to measure (the shared-support disagreement) and where to allocate (the rising transitions) matter more than how finely the budget is refined afterward.

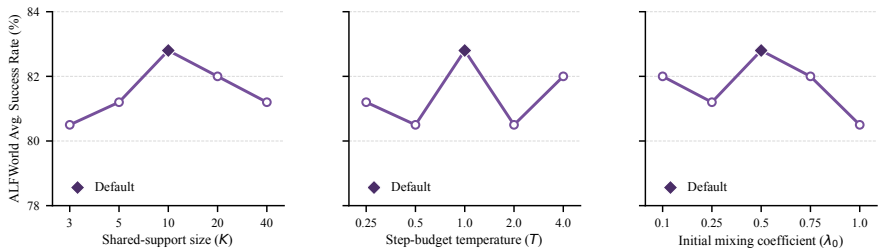


Figure 3: Hyperparameter sensitivity of TAAD with Qwen2.5-3B-Instruct on ALFWorld, varying the top-K support size K, the step-budget temperature T, and the initial mixing coefficient λ_0.

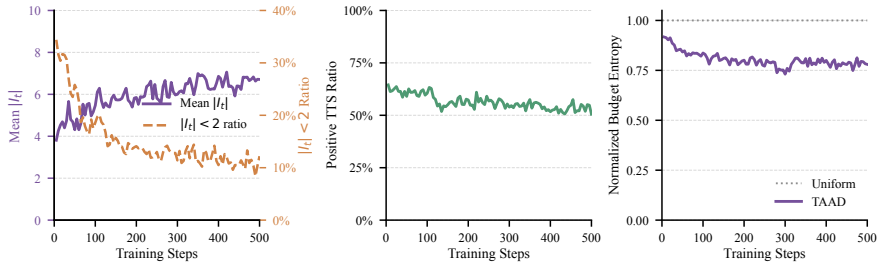


Figure 4: Mechanism analysis of TAAD with Qwen2.5-3B-Instruct on WebShop. Left: mean intersection size and fraction of near-degenerate intersections. Middle: fraction of steps with a positive TTS. Right: normalized step-budget entropy, from one (uniform) to zero (fully concentrated).

5.5 Hyperparameter Sensitivity

Figure 3 examines the sensitivity of TAAD to its three principal hyperparameters, varying one at a time. Performance is stable across all three grids: the default configuration attains the strongest result, and every alternative remains within a narrow band of it. Each sweep admits a mechanism-level reading. The top-K support size K shows a broad peak around the default, as an overly small support discards candidates the teacher endorses while an overly large one reintroduces the long-tail noise the intersection is designed to filter. The temperature T interpolates between concentrated and near-uniform budget allocation, and the moderate default balances the two regimes. The initial mixing coefficient λ_0 is similarly forgiving: even full reweighting remains competitive, consistent with the clipping bound that keeps the reshaped advantage anchored to the sequence-level signal.

5.6 Mechanism Analysis

Figure 4 examines whether the core mechanisms of TAAD remain active and non-degenerate throughout training. The shared support strengthens: the mean intersection size grows while near-degenerate intersections become rare, so the teachability signal is computed over an increasingly meaningful candidate set rather than collapsing toward the trivial case. Transition prioritization stays engaged, with roughly half the steps receiving a positive TTS, confirming that rising-disagreement transitions persist as the two policies converge. The step budget becomes more selective, as the normalized entropy declines from near-uniform allocation, yet remains far from single-step concentration, indicating that TAAD sharpens its allocation while preserving coverage elsewhere.

6 Conclusion

In this work, we identify a structural limitation of multi-step agentic reinforcement learning: a sparse trajectory-level outcome broadcasts identical credit to every decision. We address it with Teachability-Aware Agentic Distillation (TAAD), a bounded two-level advantage reweighting framework that combines teacher-mass-weighted disagreement over shared top-K support, action-surprisal-gated rising transitions, and maximum-entropy step budgeting, while preserving the sign of the sequence-level GRPO advantage. Across ALFWorld, Search-QA, and WebShop, TAAD consistently achieves the strongest aggregate results, suggesting that bounded, transition-aware credit reallocation is a promising direction for training multi-step language-model agents.

AI USE STATEMENT

Generative AI tools assisted with literature organization, language editing, and preparation of experimental layouts. The authors reviewed all AI-assisted material and take full responsibility for the content of this paper, including the validity of its claims, the accuracy of its citations, and the consistency of the reported results with the completed experiments.

ETHICS STATEMENT

This work studies training-time credit assignment for language-model agents on established, fully simulated and text-based benchmarks. It does not involve human subjects or private user data. All datasets and environments used in this paper are publicly released for research purposes, and we use them in accordance with their intended research scope and licenses. Any deployment of the trained models in real-world systems would require careful task-specific safety evaluation, access controls, and monitoring, and is therefore outside the scope of this paper.

REPRODUCIBILITY STATEMENT

Section 5 specifies the evaluation protocol, benchmark configurations, and principal hyperparameters, and Appendix B provides complete proofs of the stated theoretical properties. The supplementary material accompanying this submission includes the training configurations, environment preprocessing, skill-retrieval settings, and scripts 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 *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

Kevin Chen, Marco F. Cusumano-Towner, Brody Huval, Aleksei Petrenko, Jackson Hamburger, Vladlen Koltun, and Philipp Krähenbühl. Reinforcement learning for long-horizon interactive LLM agents. *CoRR*, abs/2502.01600, 2025. doi: 10.48550/ARXIV.2502.01600. URL https://doi.org/10.48550/arXiv.2502.01600.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, Jiarui Yuan, Huayu Chen, Kaiyan Zhang, Xingtai Lv, Shuo Wang, Yuan Yao, Xu Han, Hao Peng, Yu Cheng, Zhiyuan Liu, Maosong Sun, Bowen Zhou, and Ning Ding. Process reinforcement through implicit rewards. *CoRR*, abs/2502.01456, 2025. doi: 10.48550/ARXIV.2502.01456. URL https://doi.org/10.48550/arXiv.2502.01456.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948, 2025. doi: 10.48550/ARXIV.2501.12948. URL https://doi.org/10.48550/arXiv.2501.12948.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for LLM agent training. *CoRR*, abs/2505.10978, 2025. doi: 10.48550/ARXIV.2505.10978. URL https://doi.org/10.48550/arXiv.2505.10978.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes. *CoRR*, abs/2603.25562, 2026. doi: 10.48550/ARXIV.2603.25562. URL https://doi.org/10.48550/arXiv.2603.25562.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=5h0qf7IBZZ.

Yinghui He, Simran Kaur, Adithya Bhaskar, Yongjin Yang, Jiarui Liu, Narutatsu Ri, Liam Fowl, Abhishek Panigrahi, Danqi Chen, and Sanjeev Arora. Self-distillation zero: Self-revision turns binary rewards into dense supervision. *CoRR*, abs/2604.12002, 2026. doi: 10.48550/ARXIV.2604.12002. URL https://doi.org/10.48550/arXiv.2604.12002.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing A multi-hop QA dataset for comprehensive evaluation of reasoning steps. In Donia Scott, Núria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pp. 6609–6625. International Committee on Computational Linguistics, 2020. doi: 10.18653/V1/2020.COLING-MAIN.580. URL https://doi.org/10.18653/v1/2020.coling-main.580.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation. *CoRR*, abs/2601.20802, 2026. doi: 10.48550/ARXIV.2601.20802. URL https://doi.org/10.48550/arXiv.2601.20802.

Ijun Jang, Je Won Yeom, Juan Yeo, Hyunggyu Lim, and Taesup Kim. Stable on-policy distillation through adaptive target reformulation. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics, ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 42217–42227. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.FINDINGS-ACL.2094. URL https://doi.org/10.18653/v1/2026.findings-acl.2094.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *CoRR*, abs/2503.09516, 2025. doi: 10.48550/ARXIV.2503.09516. URL https://doi.org/10.48550/arXiv.2503.09516.

Woogyeol Jin, Taywon Min, Yongjin Yang, Swanand Ravindra Kadhe, Yi Zhou, Dennis Wei, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. *CoRR*, abs/2603.07079, 2026. doi: 10.48550/ARXIV.2603.07079. URL https://doi.org/10.48550/arXiv.2603.07079.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pp. 1601–1611. Association for Computational Linguistics, 2017. doi: 10.18653/V1/P17-1147. URL https://doi.org/10.18653/v1/P17-1147.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466, 2019. doi: 10.1162/TACL_A_00276. URL https://doi.org/10.1162/tacl_a_00276.

Ziheng Li, Liu Kang, Feng Xiao, Luxi Xing, Qingyi Si, Zhuoran Li, Weikang Gong, Deqing Yang, Yanghua Xiao, and Hongcheng Guo. Outcome-grounded advantage reshaping for fine-grained credit assignment in mathematical reasoning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 24681–24693. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.ACL-LONG.1132. URL https://doi.org/10.18653/v1/2026.acl-long.1132.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria,*

May 7-11, 2024. OpenReview.net, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Zhengxi Lu, Zhiyuan Yao, Zhu Han, Zihan Wang, Jinyang Wu, Qi Gu, Xunliang Cai, Weiming Lu, Jun Xiao, Yueting Zhuang, and Yongliang Shen. Self-distilled agentic reinforcement learning. *CoRR*, abs/2605.15155, 2026. doi: 10.48550/ARXIV.2605.15155. URL https://doi.org/10.48550/arXiv.2605.15155.

Xufang Luo, Yuge Zhang, Zhiyuan He, Zilong Wang, Siyun Zhao, Dongsheng Li, Luna K. Qiu, and Yuqing Yang. Agent lightning: Train ANY AI agents with reinforcement learning. *CoRR*, abs/2508.03680, 2025. doi: 10.48550/ARXIV.2508.03680. URL https://doi.org/10.48550/arXiv.2508.03680.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 9802–9822. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.546. URL https://doi.org/10.18653/v1/2023.acl-long.546.

Emiliano Penaloza, Dheeraj Vattikonda, Nicolas Gontier, Alexandre Lacoste, Laurent Charlin, and Massimo Caccia. Privileged information distillation for language models. *CoRR*, abs/2602.04942, 2026. doi: 10.48550/ARXIV.2602.04942. URL https://doi.org/10.48550/arXiv.2602.04942.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, volume EMNLP 2023 of *Findings of ACL*, pp. 5687–5711. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-EMNLP.378. URL https://doi.org/10.18653/v1/2023.findings-emnlp.378.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024. doi: 10.48550/ARXIV.2402.03300. URL https://doi.org/10.48550/arXiv.2402.03300.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. *CoRR*, abs/2601.19897, 2026. doi: 10.48550/ARXIV.2601.19897. URL https://doi.org/10.48550/arXiv.2601.19897.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew J. Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. *CoRR*, abs/2010.03768, 2020. URL https://arxiv.org/abs/2010.03768.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Musique: Multihop questions via single-hop question composition. *Trans. Assoc. Comput. Linguistics*, 10: 539–554, 2022. doi: 10.1162/TACL_A_00475. URL https://doi.org/10.1162/tacl_a_00475.

Hao Wang, Guozhi Wang, Han Xiao, Yufeng Zhou, Yue Pan, Jichao Wang, Ke Xu, Yafei Wen, Xiaohu Ruan, Xiaoxin Chen, and Honggang Qi. Skill-sd: Skill-conditioned self-distillation for multi-turn LLM agents. *CoRR*, abs/2604.10674, 2026a. doi: 10.48550/ARXIV.2604.10674. URL https://doi.org/10.48550/arXiv.2604.10674.

Jiaqi Wang, Wenhao Zhang, Weijie Shi, Yaliang Li, and James Cheng. TCOD: exploring temporal curriculum in on-policy distillation for multi-turn autonomous agents. *CoRR*, abs/2604.24005, 2026b. doi: 10.48550/ARXIV.2604.24005. URL https://doi.org/10.48550/arXiv.2604.24005.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *CoRR*, abs/2212.03533, 2022. doi: 10.48550/ARXIV.2212.03533. URL https://doi.org/10.48550/arXiv.2212.03533.

Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Xing Jin, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, Eli Gottlieb, Yiping Lu, Kyunghyun Cho, Jiajun Wu, Li Fei-Fei, Lijuan Wang, Yejin Choi, and Manling Li. RAGEN: understanding self-evolution in LLM agents via multi-turn reinforcement learning. *CoRR*, abs/2504.20073, 2025. doi: 10.48550/ARXIV.2504.20073. URL https://doi.org/10.48550/arXiv.2504.20073.

Peng Xia, Jianwen Chen, Hanyang Wang, Jiaqi Liu, Kaide Zeng, Yu Wang, Siwei Han, Yiyang Zhou, Xujiang Zhao, Haifeng Chen, Zeyu Zheng, Cihang Xie, and Huaxiu Yao. Skillrl: Evolving agents via recursive skill-augmented reinforcement learning. *CoRR*, abs/2602.08234, 2026. doi: 10.48550/ARXIV.2602.08234. URL https://doi.org/10.48550/arXiv.2602.08234.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. TIP: token importance in on-policy distillation. *CoRR*, abs/2604.14084, 2026. doi: 10.48550/ARXIV.2604.14084. URL https://doi.org/10.48550/arXiv.2604.14084.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled RLVR. *CoRR*, abs/2604.03128, 2026a. doi: 10.48550/ARXIV.2604.03128. URL https://doi.org/10.48550/arXiv.2604.03128.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation. *CoRR*, abs/2602.12125, 2026b. doi: 10.48550/ARXIV.2602.12125. URL https://doi.org/10.48550/arXiv.2602.12125.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pp. 2369–2380. Association for Computational Linguistics, 2018. doi: 10.18653/V1/D18-1259. URL https://doi.org/10.18653/v1/d18-1259.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. *CoRR*, abs/2207.01206, 2022. doi: 10.48550/ARXIV.2207.01206. URL https://doi.org/10.48550/arXiv.2207.01206.

Tianzhu Ye, Li Dong, Zewen Chi, Xun Wu, Shaohan Huang, and Furu Wei. Black-box on-policy distillation of large language models. *CoRR*, abs/2511.10643, 2025. doi: 10.48550/ARXIV.2511.10643. URL https://doi.org/10.48550/arXiv.2511.10643.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. *CoRR*, abs/2602.12275, 2026. doi: 10.48550/ARXIV.2602.12275. URL https://doi.org/10.48550/arXiv.2602.12275.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: an open-source LLM reinforcement learning system at scale. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5,*

2025, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/a4277440d50f1f15d2cb4c14f7e0c0d2-Abstract-Conference.html.

Yanfei Zhang, Xu Lin, and Chenglin Wu. Stepopsd: Step-aware online preference distillation for agent reinforcement learning. *CoRR*, abs/2605.27140, 2026. doi: 10.48550/ARXIV.2605.27140. URL https://doi.org/10.48550/arXiv.2605.27140.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *CoRR*, abs/2601.18734, 2026. doi: 10.48550/ARXIV.2601.18734. URL https://doi.org/10.48550/arXiv.2601.18734.

Qiyong Zhong, Mao Zheng, Mingyang Song, Xin Lin, Jie Sun, Houcheng Jiang, Xiang Wang, and Junfeng Fang. SOD: step-wise on-policy distillation for small language model agents. *CoRR*, abs/2605.07725, 2026. doi: 10.48550/ARXIV.2605.07725. URL https://doi.org/10.48550/arXiv.2605.07725.

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. Archer: Training language model agents via hierarchical multi-turn RL. *CoRR*, abs/2402.19446, 2024. doi: 10.48550/ARXIV.2402.19446. URL https://doi.org/10.48550/arXiv.2402.19446.

A Theoretical Properties

We summarize the properties of TAAD and defer proofs to Appendix B. For $\varepsilon_w < 1$, $w_{s,t}^{\mathrm{clip}} \in [1-\varepsilon_w, 1+\varepsilon_w] \subset (0,\infty)$. The reshaped advantage therefore preserves the sign of A_{seq} at each valid token, while teacher statistics change only its magnitude. Its deviation satisfies $|\hat{A}_t - A_{\mathrm{seq}}| \le \lambda\varepsilon_w|A_{\mathrm{seq}}|$. By linearity of expectation, averaging over the valid response tokens of a single trajectory, within which A_{seq} is constant, gives $\mathbb{E}_t[\hat{A}_t] = A_{\mathrm{seq}}[1+\lambda(\mathbb{E}_t[w_{s,t}^{\mathrm{clip}}]-1)]$, and the triangle inequality with the clipping bound gives the same deviation $|\mathbb{E}_t[\hat{A}_t] - A_{\mathrm{seq}}| \le \lambda\varepsilon_w|A_{\mathrm{seq}}|$. When the pre-clipped weights are approximately centered, the residual term is a small mean shift rather than a systematic bias. The $(1-\lambda)$ term is a constant GRPO anchor and does not compensate for clipping asymmetry. TAAD is therefore a bounded advantage reweighting.

Proposition 1 (Max-entropy transition-priority allocation)**.** *For fixed transition-priority scores q_s and step-budget temperature $T > 0$, solving the entropy-regularized utility maximization*

$$\max_{b\in\Delta^N} \sum_s b_s q_s - T\sum_s b_s \log b_s \tag{16}$$

yields the unique solution $b_s^\star = \mathrm{softmax}(q_s/T)$. Thus T explicitly trades off allocating budget to high-priority transitions and preserving entropy over steps on the simplex.

Proposition 2 (Condition for equivalence with importance allocation)**.** *An importance allocation $b_s \propto J_s$, including the uniform GRPO allocation ($J_s \equiv 1$), coincides with the optimizer of Proposition 1 only when its normalized distribution equals* $\mathrm{softmax}(q_s/T)$*. In particular, uniform allocation coincides with it only when all q_s are equal.*

The construction admits a limited teaching interpretation. D_t measures disagreement after renormalization on shared support, C_t weights it by the teacher probability mass on that support, and their product defines the conservative proxy M_t. The action-surprisal gate and TTS_s then form a heuristic transition-priority score. This interpretation motivates the allocation but does not identify M_t with teacher correctness, eliminable uncertainty, or exact expected teaching information gain.

B Proofs

This section proves the statements in Appendix A. Throughout, let $T > 0$, $0 \le \lambda \le 1$, and $0 \le \varepsilon_w < 1$, as required by the entropy-regularized allocation and the convex advantage mixing rule.

B.1 Proof of Proposition 1

Proof. Consider the objective

$$F(b) = \sum_{s=1}^{N} b_s q_s - T \sum_{s=1}^{N} b_s \log b_s \tag{17}$$

on the probability simplex Δ^N. Because $T > 0$, the entropy term is strictly concave on the interior of the simplex. Moreover, the derivative of the entropy barrier $-T\, b_s \log b_s$ diverges to $+\infty$ as any b_s approaches zero from above. The unique maximizer therefore lies in the interior, where the equality constraint $\sum_s b_s = 1$ is the only active constraint.

Introducing a Lagrange multiplier ν for this constraint gives

$$\mathcal{J}(b, \nu) = \sum_{s=1}^{N} b_s q_s - T \sum_{s=1}^{N} b_s \log b_s + \nu \left(\sum_{s=1}^{N} b_s - 1 \right). \tag{18}$$

The stationary condition for each s is

$$\frac{\partial \mathcal{J}}{\partial b_s} = q_s - T(\log b_s + 1) + \nu = 0. \tag{19}$$

Thus $\log b_s = (q_s + \nu - T)/T$, so $b_s = c \exp(q_s/T)$ for a constant c independent of s. Enforcing $\sum_s b_s = 1$ yields

$$b_s^\star = \frac{\exp(q_s/T)}{\sum_{j=1}^{N} \exp(q_j/T)} = \mathrm{softmax}(q_s/T). \tag{20}$$

Strict concavity establishes uniqueness.

B.2 Proof of Proposition 2

Proof. Let $J_s \geq 0$ be importance scores that are not all zero, and let the corresponding normalized importance allocation be $b_s^J = J_s / \sum_j J_j$. Proposition 1 has the unique optimizer $b_s^\star = \mathrm{softmax}(q_s/T)$. Hence the two allocations coincide if and only if

$$\frac{J_s}{\sum_j J_j} = \frac{\exp(q_s/T)}{\sum_j \exp(q_j/T)} \tag{21}$$

for every s, which is precisely the stated condition.

For uniform allocation, $J_s \equiv 1$ and $b_s^J = 1/N$. This equals $\mathrm{softmax}(q_s/T)$ if and only if all exponentials $\exp(q_s/T)$ are equal. Since the exponential function is injective and $T > 0$, this holds if and only if all q_s are equal.

B.3 Proof of the Bounded Reweighting Properties

Proof of per-token sign preservation. By construction,

$$w_{s,t}^{\mathrm{clip}} \in [1 - \varepsilon_w, 1 + \varepsilon_w]. \tag{22}$$

Since $0 \leq \varepsilon_w < 1$, this interval is strictly positive. Define the advantage multiplier

$$g_{s,t} = (1 - \lambda) + \lambda w_{s,t}^{\mathrm{clip}}. \tag{23}$$

Because $g_{s,t}$ is a convex combination of 1 and the positive quantity $w_{s,t}^{\mathrm{clip}}$, it is strictly positive. Therefore $\hat{A}_t = A_{\mathrm{seq}} g_{s,t}$ has the same sign as A_{seq}, including the case $A_{\mathrm{seq}} = 0$.

Proof of the per-token deviation bound. Using the definition of $\hat{A}_t$,

$$\hat{A}_t - A_{\mathrm{seq}} = A_{\mathrm{seq}} \lambda \left(w_{s,t}^{\mathrm{clip}} - 1 \right). \tag{24}$$

The clipping interval gives $|w_{s,t}^{\mathrm{clip}} - 1| \leq \varepsilon_w$. Taking absolute values proves

$$\left| \hat{A}_t - A_{\mathrm{seq}} \right| \leq \lambda \varepsilon_w \left| A_{\mathrm{seq}} \right|. \tag{25}$$

Table 3: Training-efficiency analysis derived from the validation curves in Figure 2. *Final* repeats the step-500 evaluation of Table 1 for the four methods shown in Figure 2. *Train Mean* averages all 100 recorded evaluations, and *Steps to 90%* reports the first training step at which the smoothed curve reaches 90% of that method's own final score. **Best** values per setting and column are highlighted.

Method	Final	Train Mean	Steps to 90%
Qwen2.5-3B-Instruct / ALFWorld			
GRPO	72.7	59.4	**220**
RLSD	74.2	57.9	320
SDAR	79.7	60.4	350
TAAD (Ours)	**82.8**	**61.2**	325
Qwen2.5-3B-Instruct / Search-QA			
GRPO	41.0	35.2	220
RLSD	41.3	35.0	**215**
SDAR	44.0	36.5	250
TAAD (Ours)	**44.8**	**37.3**	250
Qwen2.5-3B-Instruct / WebShop			
GRPO	61.7	45.6	**335**
RLSD	66.4	47.7	345
SDAR	63.3	44.8	355
TAAD (Ours)	**72.7**	**51.1**	360
Qwen2.5-7B-Instruct / ALFWorld			
GRPO	75.8	60.9	**200**
RLSD	75.0	58.8	320
SDAR	82.0	62.0	320
TAAD (Ours)	**85.9**	**64.0**	330
Qwen2.5-7B-Instruct / Search-QA			
GRPO	43.6	37.8	**190**
RLSD	43.4	37.4	215
SDAR	44.6	37.6	220
TAAD (Ours)	**47.3**	**39.0**	310
Qwen2.5-7B-Instruct / WebShop			
GRPO	68.0	50.6	**295**
RLSD	70.3	52.0	300
SDAR	71.1	51.5	305
TAAD (Ours)	**78.1**	**54.4**	310

Proof of the mean-scale identity and bound. Taking the expectation over any fixed collection of valid response tokens from the same trajectory and using that A_{seq} is constant within that trajectory gives

$$\mathbb{E}_t[\hat{A}_t] = A_{\text{seq}}\left[(1-\lambda) + \lambda \mathbb{E}_t[w_{s,t}^{\text{clip}}]\right] = A_{\text{seq}}\left[1 + \lambda\left(\mathbb{E}_t[w_{s,t}^{\text{clip}}] - 1\right)\right]. \tag{26}$$

The identity holds without a centering assumption. If the pre-clipped weights are approximately centered, this expression characterizes the remaining small mean shift. Since $|w_{s,t}^{\text{clip}} - 1| \le \varepsilon_w$ and $|\mathbb{E}[X]| \le \mathbb{E}|X|$, the triangle inequality yields

$$\left|\mathbb{E}_t[\hat{A}_t] - A_{\text{seq}}\right| \le \lambda \varepsilon_w \, |A_{\text{seq}}|. \tag{27}$$

C Additional Experimental Analysis

Table 3 quantifies the training dynamics of Figure 2. TAAD attains the highest training mean in all six settings, confirming that its advantage accumulates over the course of optimization rather than emerging only at the final evaluation. The steps-to-90% column shows that GRPO saturates earliest but at a lower level. TAAD reaches 90% of a strictly higher final score within steps comparable to the strongest distillation baselines in five of the six settings; the later saturation on 7B Search-QA is consistent with its margin still growing at the end of training, as observed in Figure 2.

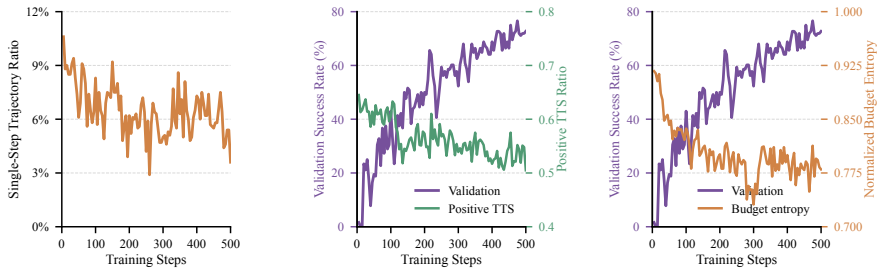


Figure 5: Extended diagnostics of TAAD with Qwen2.5-3B-Instruct on WebShop. Left: fraction of single-step trajectories excluded from the normalized budget entropy. Middle: validation success rate with the positive-TTS ratio. Right: validation success rate with normalized budget entropy.

Figure 5 extends the mechanism analysis of Figure 4 on the same training run. The left panel reports the entropy statistic's exclusion basis: the fraction of single-step trajectories stays between roughly 3% and 11% throughout training, so the normalized entropy is computed over most trajectories. The middle and right panels align the mechanism signals with the validation performance curve. The positive-TTS ratio declines mildly (from about 0.65 to 0.50) and the normalized budget entropy from about 0.92 to 0.78 while the success rate rises, indicating that the growing selectivity of the step budget co-evolves with policy competence rather than a collapse of the teachability signal.