SCOUT: STATE-CONDITIONAL COUNTERFACTUAL PRIVILEGE DISTILLATION FOR AGENTIC REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy self-distillation gives a teacher privileged task information while training the student on its own interaction histories. Yet a correct solution does not need describe how to proceed after the student makes a mistake. We introduce **S**tate-conditional **COU**nterfactual privilege dis**T**illation (**SCOUT**), which adapts privileged guidance to the student's current state. A frozen teacher scores the same student history and token prefix under three views: a successful task plan (Gold View), reusable skills with preconditions and recovery steps (Skill View), and action structure with instance-specific privileged content removed (Neutral View). An online buffer supplies skills from successful training trajectories. A parameter-free gate uses environment-specific compatibility and applicability signals to mix the three teacher distributions into one distillation target. All additional components are training-only and inference uses the student alone. Across three backbones on ALFWorld, WebShop, and search-based question answering, SCOUT improves aggregate performance over Gold-only distillation both alone and with Group Relative Policy Optimization (GRPO). With Qwen3.5-4B and GRPO, it raises ALFWorld success from 88.2% to 93.2%, WebShop success from 68.7% to 77.4%, and search-QA exact match from 45.74% to 49.93%. Ablations on Qwen3.5-4B support adaptive weighting and retaining all three views, although gains vary across task categories.

1 INTRODUCTION

Large language model (LLM) agents operate through multi-turn interactions in which each action changes the observations available for subsequent decisions (Yao et al., 2023; Huang et al., 2023; Schick et al., 2023; Wei et al., 2022; Gao et al., 2023). To complete a task, an agent needs to choose actions that are valid in the current state, interpret the resulting feedback, and revise its plan when earlier decisions lead away from the intended route. To make the agent more capable, it needs to learn more than a single successful path to the goal. It also needs to know how to keep going from the intermediate states that its own actions produce. Offline demonstrations provide explicit action sequences, but they can not cover all states induced by the learner's own behavior. As a result, errors may compound in subsequent decisions (Ross et al., 2011). Outcome-based reinforcement learning directly optimizes task success (Shao et al., 2024; Schulman et al., 2017; Ouyang et al., 2022; Rafailov et al., 2023; DeepSeek-AI, 2025; Yu et al., 2025; Liu et al., 2025; Zhang et al., 2026a), but a trajectory-level reward provides limited information about which individual tokens or actions should be corrected. Knowledge distillation provides token-level supervision by transferring a teacher's predictive distribution to the student (Hinton et al., 2015), and sequence-level distillation extends this principle to structured generation (Kim & Rush, 2016). The remaining challenge for multi-turn tasks is to deliver such fine-grained guidance on the histories visited by the student.

On-policy distillation addresses this question by moving teacher supervision onto the learner's own distribution. Rather than training only on offline teacher outputs, it samples student-generated sequences and evaluates them with a teacher at the prefixes the student visits (Agarwal et al., 2024). On-policy self-distillation (OPSD) further uses a single model as both the student and the teacher. The student acts without any privileged information, while the teacher gets extra task information, such as a verified solution. Matching their token distributions can internalize this information while

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.

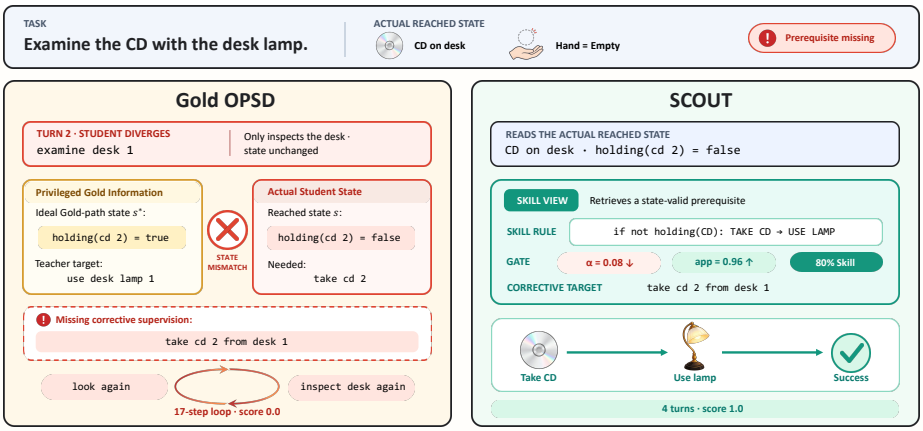

Figure 1: **Skill guidance addresses the state mismatch left by a fixed Gold path. Left:** The Gold path assumes the CD is held, leaving the empty-handed student without suitable recovery guidance. **Right:** The Skill view supplies a recovery plan: retrieve the CD, then use the lamp. View annotations illustrate training-time supervision.

keeping inference student-only (Zhao et al., 2026). Recent work has adapted the intervention schedule to multi-turn interaction: SAGE-OPD selects turns using environment feedback and teacher judgment, while TurnOPD allocates rollout depth and balances supervision across turns (Zhou et al., 2026a;b). These methods better align teacher feedback with where and when the student needs training. They leave a separate issue unresolved. When the teacher is conditioned on a task-level solution, is that privileged context still suitable for the state produced by the student's preceding actions?

Our motivating observation is that even when the student and teacher share the same history, the privileged context given to the teacher may no longer apply to the student's current state. Consider an agent asked to examine a CD with a desk lamp (Figure 1). Using the lamp requires the agent to hold the CD, but a student that repeatedly inspects the desk has not met this precondition and must first take the CD. A plan that would complete the task might still contain this exact action, yet it may not show how to recover from the specific mistakes the student has already made. The teacher still evaluates the student's actual history. The problem is that the privileged context handed to the teacher is fixed and doesn't adapt to that history. This distinction separates two things. *Task-level correctness* asks whether a plan reaches the goal along its intended route. *State-level applicability* asks whether the same information still gives useful guidance from the state the student has actually reached. We therefore ask whether adapting privileged information to the reached state improves supervision, compared with using one fixed task plan throughout the interaction.

To alleviate these issues, we propose **S**tate-conditional **COU**nterfactual privilege dis**T**illation (**SCOUT**), which adapts the teacher's privileged context to the student's current state. At each turn, the same frozen teacher scores the same student history and sampled token prefix under three views. The Gold view supplies a successful task plan, and the Skill view supplies a reusable skill skeleton with preconditions and recovery steps, and the Neutral view preserves the action structure while removing instance-specific privileged content. Skills are drawn from an online buffer built from successful training trajectories. A parameter-free gate uses environment-specific state signals to assess Gold compatibility and skill applicability, then combines the three teacher distributions into one probability-mixture target. The target retains compatible Gold guidance, shifts weight toward applicable Skill guidance, and assigns the remainder to Neutral. SCOUT changes the teacher target rather than introducing view-specific losses, supports pure distillation and auxiliary distillation with GRPO, and uses only the student policy at inference.

We evaluate SCOUT on ALFWorld (Shridhar et al., 2021), WebShop (Yao et al., 2022), and a seven-dataset search-based question-answering suite, using Qwen3-4B, Qwen3.5-4B, and Qwen3.5-9B. SCOUT improves over Gold-only OPSD on every benchmark aggregate for all three backbones, with and without GRPO. With Qwen3.5-4B and GRPO, it raises ALFWorld success from 88.2% to 93.2%, WebShop success from 68.7% to 77.4%, and search-QA exact match from 45.74% to 49.93%. Ablations on this backbone yield lower ALFWorld and search-QA results when state-dependent weights are replaced with fixed weights or any one of the three views is removed.

Our contributions are as follows:

- We identify a state mismatch in on-policy self-distillation: a task-correct privileged plan may no longer provide useful teacher guidance after the student reaches an off-route state. We distinguish task-level correctness from state-level applicability and formulate supervision on the student's reached state.
- We introduce SCOUT, a state-conditional counterfactual privilege distillation method. A frozen teacher evaluates the same student history and sampled prefix under Gold, Skill, and Neutral contexts, and a parameter-free gate mixes the resulting teacher probabilities into one target using Gold compatibility and skill applicability.
- We evaluate SCOUT across three backbones on ALFWorld, WebShop, and search-based QA, under both pure distillation and GRPO-augmented training. We compare SCOUT with methods that combine distillation and outcome-based reinforcement learning, and SCOUT+GRPO achieves the best aggregate performance across all backbones and benchmarks. Ablations show that all three views and state-dependent weighting contribute.

2 Related Work

On-policy distillation for multi-turn agents. Knowledge distillation transfers a teacher's predictive distribution to a student, and sequence-level distillation extends this idea to structured generation (Hinton et al., 2015; Kim & Rush, 2016). Unlike offline distillation, interactive imitation learning and generalized knowledge distillation collect supervision on learner-visited states (Ross et al., 2011; Agarwal et al., 2024). For multi-turn agents, SAGE-OPD selects intervention turns using environment feedback and teacher judgment, whereas TurnOPD allocates rollout depth and balances supervision across turns (Zhou et al., 2026a;b). These methods vary supervision across turns. SCOUT instead changes the teacher target at a visited state by mixing distributions from different privileged views while keeping the student-generated history fixed.

Privileged context in self-distillation. Learning using privileged information studies supervision available during training but absent at deployment, and generalized distillation connects this setting to teacher–student transfer (Lopez-Paz et al., 2015). OPSD gives the teacher verified solutions on student-generated prefixes (Zhao et al., 2026). CANON uses a sampled majority-answer solution as privileged context, while LOPD learns continuous latent context from retrieved experiences (Gkountouras et al., 2026; Zhang et al., 2026b). In contrast, SCOUT uses explicit Gold, Skill, and Neutral views, whose weights are determined analytically from Gold compatibility and skill applicability.

Self-distillation within agentic reinforcement learning. SDAR gates an auxiliary distillation loss with detached token-level signals, and OCSD uses the score difference between replay views with and without future observations to modulate token-level updates (Lu et al., 2026; Yang et al., 2026b). Both methods adjust the RL update while leaving the teacher target unchanged. SCOUT instead reweights the Gold, Skill, and Neutral teacher distributions before distillation.

Skills and recovery from experience. Inner Monologue incorporates environment feedback into planning, whereas Reflexion stores verbal reflections for later trials (Huang et al., 2023; Shinn et al., 2023). ExpeL extracts reusable insights from experience, and Voyager maintains an executable skill library for lifelong exploration (Zhao et al., 2024; Wang et al., 2024). These methods retain an external experience mechanism at deployment. Skill-SD summarizes successful trajectories into teacher-only skills, and OPID converts hindsight skills into a token-level RL advantage (Wang et al., 2026; Yang et al., 2026a). SCOUT also avoids test-time skill retrieval, but uses skill preconditions and re-entry steps to construct one component of its teacher target. Its analytic gate mixes this component with Gold and Neutral views rather than routing skills through a token-level advantage.

3 Method

A fixed privileged plan may no longer provide useful supervision after the student reaches a different state. We address this mismatch with SCOUT, which constructs a state-dependent teacher target for

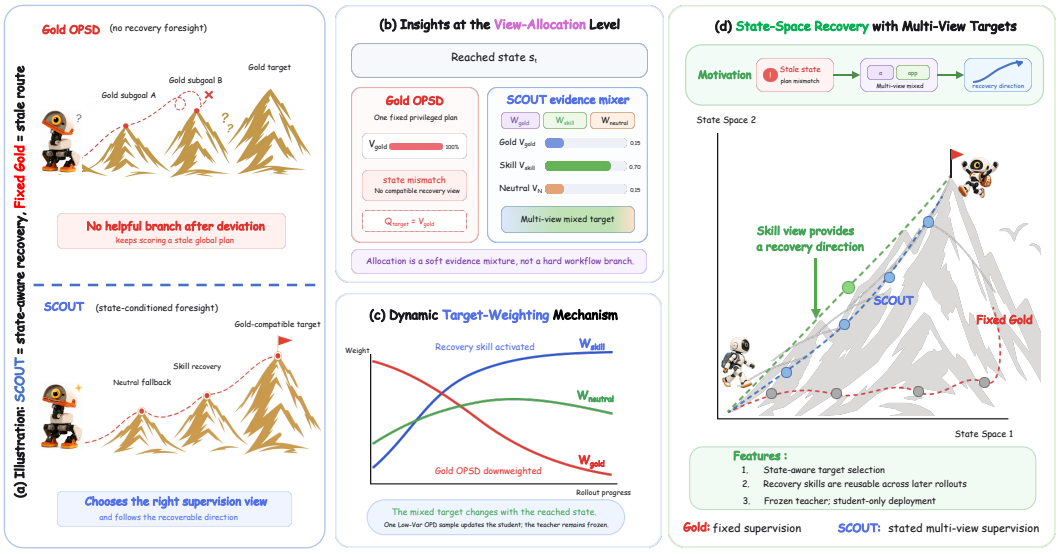

Figure 2: **SCOUT combines teacher-view construction with state-conditioned distillation. View construction:** a frozen teacher scores the same student history and token prefix under Gold, Skill, and Neutral contexts, giving a task plan, reusable recovery guidance from an online buffer of successful trajectories, and action structure free of instance-specific privileged content. **State-conditioned distillation:** a parameter-free gate weights the three views by Gold compatibility and skill applicability, raising recovery guidance as the plan stops applying.

tokens generated by the student. The same frozen teacher evaluates the student history under three privileged contexts, and a deterministic gate mixes the resulting probabilities. Figure 2 summarizes the training process, and Figure 3 illustrates the three contexts for a fixed student history.

3.1 Student Histories and Teacher Views

Let x denote a task and $h_t = (x, o_0, y_0, \ldots, y_{t-1}, o_t)$ the interaction history before turn t, where o_t is the current observation. At training iteration k, the rollout policy π_{θ_k} generates a response $y_t = (y_{t,1}, \ldots, y_{t,L_t})$. The environment executes the action parsed from this response and returns the next observation. For a token u in vocabulary $\mathcal{Y}$, define

$$c_{t,l} = (h_t, y_{t,<l}), \qquad p_{\theta,t,l}(u) = \pi_\theta(u \mid c_{t,l}), \qquad y_{t,l} \sim p_{\theta_k,t,l}. \tag{1}$$

The response may contain reasoning and action-format tokens in addition to the executable action. A loss mask $m_{t,l}$ identifies generated tokens used for training. Environment observations are not prediction targets. We write $s_t = f_e(h_t)$ for the state summary extracted from the history in environment e. This summary need not be a complete or Markov representation of the simulator state.

A frozen teacher with parameters ϕ evaluates three contexts z_t^v, indexed by $v \in \mathcal{V} = \{G, S, N\}$:

$$q_{t,l}^v(u) = \pi_\phi(u \mid h_t, z_t^v, y_{t,<l}), \qquad v \in \{G, S, N\}. \tag{2}$$

All views retain the same student history and sampled prefix. Only the additional privileged context changes. *Counterfactual views* are alternative teacher-conditioning contexts for the same history.

Gold view. The context $z_t^G = z^G(x)$ contains a successful plan for the task. It supplies the global task route, but need not describe every recovery step after a student mistake. Its token distribution still changes with h_t and $y_{t,<l}$, even though the privileged plan remains fixed.

Skill view. The context z_t^S contains a reusable skill retrieved from the training-time buffer. A skill specifies preconditions, an action schema, expected effects, and recovery or re-entry rules. These rules describe how to resume an unfinished subgoal from a compatible intermediate state.

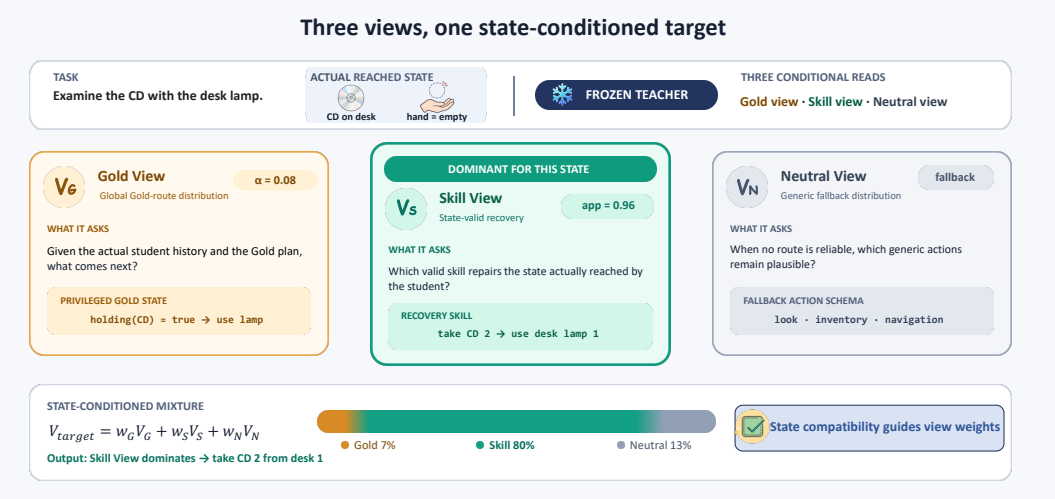

Figure 3: **Three teacher contexts for the same student history.** Gold supplies a successful task plan, Skill supplies preconditions and recovery steps, and Neutral preserves action structure without instance-specific privileged content. All views retain the actual history and sampled token prefix; the weights shown are illustrative, with the implemented mixture defined in Eq. 7.

Neutral view. The context z_t^N retains the action structure of the privileged plan while removing instance-specific entities and solution content. It provides a structural fallback when Gold and Skill receive little weight.

3.2 Online Skills and State Applicability

Let $\mathcal{B}_k$ be the skill buffer available at iteration k. We represent a skill as $\kappa = (P_\kappa, A_\kappa, E_\kappa, R_\kappa)$, where the entries denote preconditions, action schema, expected effects, and recovery/re-entry rules. Retrieval uses the task and current state to select relevant skill guidance:

$$\kappa_t = \mathrm{Retrieve}_e(\mathcal{B}_k, x, s_t), \qquad z_t^S = \mathrm{Render}(\kappa_t). \tag{3}$$

For example, a cleaning skill can require the target object to be held before cleaning, and include a re-entry rule for acquiring the correct object after a mistaken pickup. These are teacher-conditioning instructions, not hard constraints on the student's actions.

The buffer starts from historical successful training trajectories and grows from new successes. A frozen teacher abstracts environment-confirmed successful interactions into skill descriptions, which are validated and stripped of instance-specific identifiers before insertion. If $\mathcal{T}_k$ is the current rollout batch, the update can be written as

$$\mathcal{K}_k = \{\mathrm{Extract}_\phi(\tau) : \tau \in \mathcal{T}_k,\ \mathrm{Success}(\tau) = 1\}, \qquad \mathcal{B}_{k+1} = \mathrm{Update}(\mathcal{B}_k, \mathrm{Validate}(\mathcal{K}_k)). \tag{4}$$

The snapshot $\mathcal{B}_k$ remains fixed while the current batch is scored. Skills extracted from the current rollout batch enter the buffer next iteration. Evaluation trajectories never update the buffer.

Two bounded signals determine how the views are used. Gold compatibility measures whether the plan's assumptions fit the current state; skill applicability measures whether the retrieved skill has a suitable entry point. We denote them by

$$\alpha_t = g_{T_g}\left(C_e(s_t, z^G)\right), \qquad \beta_t = A_e(s_t, \kappa_t), \qquad \alpha_t, \beta_t \in [0, 1]. \tag{5}$$

Here C_e and A_e are environment-specific extractors. The endpoint-preserving sigmoid map g_{T_g} smooths compatibility with temperature T_g, while satisfying $g_{T_g}(0) = 0$ and $g_{T_g}(1) = 1$. In ALFWorld, the signals use object identity and task-precondition information, including inventory and transformation status. The semantic roles of the signals and the mixing rule are shared across environments, but their extraction rules are not learned or assumed to transfer unchanged. If no skill is retrieved, we set its applicability to zero.

3.3 A Hierarchical State-Conditional Gate

The gate first allocates weight according to Gold compatibility and then uses skill applicability to divide the remaining guidance. With a neutral-reserve coefficient $\eta \in [0, 1]$, let $b_t = 1 - \eta(1 - \alpha_t)$. The weights are

$$w_t^G = b_t \alpha_t, \qquad w_t^S = b_t(1 - \alpha_t)\beta_t, \qquad w_t^N = 1 - w_t^G - w_t^S. \tag{6}$$

There are no trainable gate parameters. Each turn's weights apply to its loss-bearing tokens, rather than being computed from the final trajectory state and broadcast backward to earlier turns.

The weights are nonnegative and sum to one. At full Gold compatibility, the target becomes Gold-only. When Gold is incompatible, applicability determines whether Skill or Neutral receives the remaining mass. Appendix D.1 gives the decomposition, boundary cases, and monotonicity.

3.4 A Single Probability-Mixture Target

SCOUT mixes probabilities before computing the distillation loss. To account for unavailable views, let $a_{t,l}^v \in \{0, 1\}$ indicate whether view v provides a valid score at prefix $c_{t,l}$. Whenever the denominator is positive, define

$$\bar{w}_{t,l}^v = \frac{a_{t,l}^v w_t^v}{\sum_{v' \in \mathcal{V}} a_{t,l}^{v'} w_t^{v'}}, \qquad q_{t,l}^\star(u) = \sum_{v \in \mathcal{V}} \bar{w}_{t,l}^v q_{t,l}^v(u). \tag{7}$$

A missing Skill score is excluded rather than treated as a zero token probability. If no positive-weight mixture can be formed, the baseline Gold score is used.

For the sampled token, the mixture is evaluated stably in log space:

$$\log q_{t,l}^\star(y_{t,l}) = \log \sum_{v:\, \bar{w}_{t,l}^v > 0} \exp\big(\log \bar{w}_{t,l}^v + \log q_{t,l}^v(y_{t,l})\big)\,. \tag{8}$$

Mixing before distillation lets one view support a token that another scores poorly, unlike separate losses that penalize disagreement with every view. Appendix D.2 formalizes this distinction.

3.5 Training Objective and Deployment

SCOUT replaces the Gold teacher probability in the baseline on-policy distillation objective (Agarwal et al., 2024; Zhao et al., 2026) with the mixed target $q^\star$ from Eq. 7. Let L_{dist} denote the resulting sampled distillation loss, computed on loss-bearing prefixes collected by the current rollout policy. Appendix D.3 defines the surrogate and relates it to the reverse-KL reference objective.

SCOUT can be trained alone or combined with the baseline GRPO loss (Shao et al., 2024):

$$L_{\text{train}} = \lambda_{\text{RL}} L_{\text{GRPO}} + \lambda_{\text{dist}} L_{\text{dist}}. \tag{9}$$

For pure distillation, we set $\lambda_{\text{RL}} = 0$. In the GRPO-augmented setting, the environment reward, group-relative advantage, and clipping rule remain unchanged from the baseline. At deployment, the teacher, privileged contexts, gate, and skill buffer are removed. The student acts only from its ordinary task and interaction history.

4 Experiments

4.1 Experimental Setup

Training setup. We report two training regimes that isolate the contribution of the teacher target. In the pure-distillation setting, the student minimizes the distillation loss alone. In the GRPO-augmented setting, the same distillation loss is added to the baseline GRPO objective (Shao et al., 2024), with the environment reward, group-relative advantage, and clipping rule unchanged. During training, a parameter-free gate combines the Gold, Skill, and Neutral views. The Skill view is drawn from an online buffer built from successful training trajectories. At evaluation, the student runs with no privileged plan, no skill retrieval, no teacher calls, and no gate. Appendix A records the full training configuration.

Table 1: **Performance comparison across ALFWorld, Search-based QA, and WebShop.** We report ALFWorld success (%), Search-QA exact match (%), and WebShop score/success on a 0–100 scale. Each model forms a separate block. The highest mean in each column is **bold** with green shading, and the second-highest distinct mean is underlined with yellow shading, including ties. OPSD denotes Gold-only OPSD. Means are shown here and complete evaluation-seed variance and ID/OOD results appear in Appendix B.

	ALFWorld							Search-based QA								WebShop	
Method	Pick	Look	Clean	Heat	Cool	Pick2	Avg	NQ	Triv	Pop	Hotp	2Wk	MuS	Bam	Avg	Score	Succ.
Qwen3-4B																	
Vanilla	69.5	32.3	22.4	17.9	8.7	24.4	31.0	29.78	49.58	30.90	26.89	30.63	7.20	25.60	33.15	14.5	4.8
OPSD	79.7	24.7	**42.5**	28.2	15.2	40.7	41.6	34.42	51.35	35.68	26.09	31.49	7.76	25.87	35.30	35.4	8.8
GRPO	89.3	51.6	32.8	43.6	30.4	**71.5**	54.0	44.02	63.91	45.68	41.66	39.54	17.17	71.20	46.21	83.1	66.0
OPSD+GRPO	88.1	53.8	31.6	53.0	26.8	59.3	52.7	43.85	64.24	46.69	**42.97**	39.58	**17.25**	**72.00**	46.75	81.6	62.8
OCSD	89.8	51.6	33.3	53.0	28.3	59.3	53.4	43.82	61.88	**47.16**	40.53	40.90	15.89	71.20	46.27	83.7	69.0
SDAR	88.7	37.6	42.0	55.6	37.0	61.8	55.6	41.02	58.43	45.69	35.34	38.55	16.55	68.80	43.63	79.2	58.1
SCOUT	79.1	32.3	41.4	39.3	15.9	40.7	43.8	39.03	60.33	36.00	33.96	35.32	12.16	31.20	39.95	40.8	28.2
SCOUT+GRPO	**94.4**	**74.2**	33.9	**68.4**	**43.5**	67.5	**63.0**	**45.43**	**64.56**	47.00	42.62	**41.00**	17.00	70.40	**47.30**	**87.0**	**75.5**
Qwen3.5-4B																	
Vanilla	79.7	6.5	15.5	5.1	8.7	39.0	29.2	30.75	50.11	32.03	25.87	30.82	7.07	38.40	33.57	14.6	8.6
OPSD	83.6	18.3	41.4	17.1	23.2	64.2	44.8	24.40	50.46	24.34	25.17	31.08	6.43	33.33	31.00	30.3	19.3
GRPO	93.8	96.8	86.8	79.5	87.7	91.1	89.2	45.40	**63.79**	44.03	40.66	40.78	15.93	70.40	45.93	82.3	64.0
OPSD+GRPO	95.5	76.3	88.5	82.9	87.7	**91.9**	88.2	45.29	60.37	45.26	44.00	39.56	16.88	**72.00**	45.74	84.5	68.7
OCSD	92.7	93.5	89.7	65.8	86.2	82.9	85.8	44.13	62.99	48.24	42.55	42.71	13.74	66.40	47.45	81.9	62.7
SDAR	94.4	73.1	**93.1**	77.8	92.0	81.3	87.0	42.85	60.29	46.35	40.58	40.47	18.45	60.00	45.63	80.1	65.4
SCOUT	85.9	18.3	44.3	23.1	32.6	62.6	48.1	40.03	59.16	38.87	38.88	39.27	11.87	60.80	42.28	51.1	33.8
SCOUT+GRPO	**96.0**	**97.8**	92.5	**87.2**	**93.5**	**91.9**	**93.2**	**46.51**	61.20	**50.50**	**46.32**	**47.77**	**20.11**	71.20	**49.93**	**88.9**	**77.4**
Qwen3.5-9B																	
Vanilla	78.0	54.8	20.7	25.6	4.3	48.8	39.1	30.97	54.85	33.73	46.32	31.83	8.77	40.00	38.35	22.9	15.2
OPSD	89.3	49.5	33.9	60.7	29.7	65.0	55.4	29.70	56.36	31.19	29.43	33.77	11.54	40.80	36.07	40.8	28.0
GRPO	**97.2**	95.7	89.7	76.9	81.2	94.3	89.4	45.98	62.37	47.54	41.61	40.31	17.38	71.20	46.72	80.2	65.5
OPSD+GRPO	93.2	93.5	83.3	77.8	84.8	94.3	87.7	46.84	65.34	50.71	46.55	45.88	19.61	72.00	50.47	78.3	65.0
OCSD	92.1	92.5	89.1	71.8	82.6	84.6	85.9	47.89	**66.99**	49.80	45.12	45.60	20.02	71.20	50.40	82.2	66.7
SDAR	96.0	79.6	97.7	**88.9**	**96.4**	**95.1**	93.4	48.06	61.61	47.70	47.00	40.36	18.99	**72.80**	47.60	83.1	67.1
SCOUT	90.4	65.6	65.5	24.8	34.1	69.9	60.5	45.37	62.87	37.05	39.32	38.07	18.78	52.80	43.04	55.5	38.2
SCOUT+GRPO	96.6	**98.9**	**98.3**	83.8	92.0	91.9	**93.9**	**48.17**	65.55	**51.80**	**47.72**	**46.66**	**21.10**	**72.80**	**51.34**	**89.1**	**77.8**

Base Models and Benchmarks. We evaluate Qwen3-4B, Qwen3.5-4B, and Qwen3.5-9B, spanning two model generations and a range of sizes. We evaluate on three established benchmarks. ALFWorld (Shridhar et al., 2021) tests embodied household interaction across six task families with different intermediate object-state transformations, and we report success for each family and their average. WebShop (Yao et al., 2022) tests web navigation; we report normalized task-completion score and exact success rate. Search-based question answering follows the Search-R1 protocol and covers seven datasets (Kwiatkowski et al., 2019; Yang et al., 2018; Joshi et al., 2017; Mallen et al., 2023; Ho et al., 2020; Trivedi et al., 2022; Press et al., 2023; Jin et al., 2025; Song et al., 2025a;b). We report exact match (EM) for each dataset and the aggregate. Success and EM are percentages, while WebShop scores use a 0–100 scale. We keep these measures separate rather than combining them into one overall score. Appendix B lists all task families and datasets.

Baselines and Evaluation. We compare against five baselines that isolate different aspects of the method. Vanilla uses the backbone with no task-specific post-training. Gold-only OPSD and SCOUT form the pure-distillation comparison: both replace the same baseline teacher with a different privileged target, while sharing the remaining objective. GRPO, Gold-only OPSD+GRPO, and SCOUT+GRPO show what reward-based learning adds to each target. OCSD (Yang et al., 2026b) and SDAR (Lu et al., 2026) are recent distillation methods that adjust RL loss without changing the teacher target. Each method contributes one checkpoint, evaluated under three inference seeds. Table 1 reports the resulting means, and Appendix B gives complete means and variance.

4.2 Main Results

SCOUT+GRPO gives the strongest aggregate performance. SCOUT+GRPO holds the highest aggregate mean in all twelve aggregate cells of Table 1, with ALFWorld success reaching 63.0,

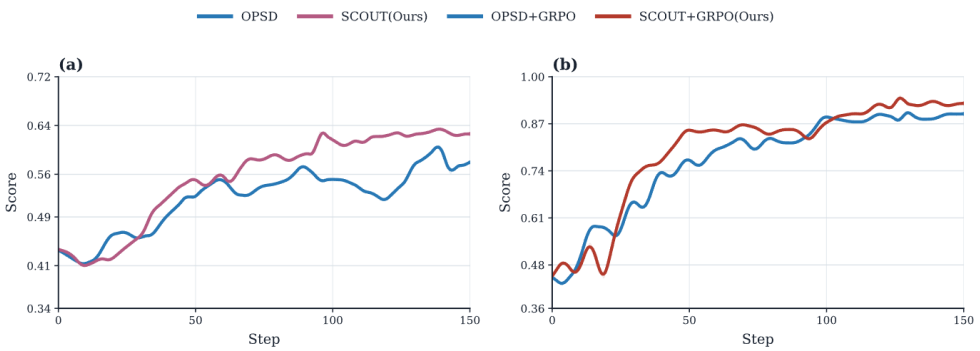

Figure 4: **ALFWorld evaluation dynamics for Qwen3.5-9B.** The left panel compares pure Gold-only OPSD with SCOUT; the right compares their GRPO-augmented variants. Curves show the recorded evaluation score at each logged step.

93.2, and 93.9, search-QA EM 47.30, 49.93, and 51.34, WebShop score 87.0, 88.9, and 89.1, and WebShop success 75.5, 77.4, and 77.8 across the three backbones, and margins over the closest competitor ranging from 0.5 points on Qwen3.5-9B ALFWorld success to 10.7 points on Qwen3.5-9B WebShop success. This advantage follows from the state-conditioned target: the gate keeps the teacher useful in states where a static Gold plan would mislead, so GRPO's reward reinforces exactly the continuations that recover from drift, compounding both signals.

SCOUT outperforms Gold-only OPSD under pure distillation. Swapping Gold-only OPSD for SCOUT raises ALFWorld success by 2.2, 3.3, and 5.1 points, WebShop score by 5.4, 20.8, and 14.7, and search-QA EM by 4.65, 11.28, and 6.97 across the three backbones. Gold-only OPSD, by contrast, leaves two of the three backbones below Vanilla on the search suite: Qwen3.5-4B drops to 31.00 EM against 33.57, and Qwen3.5-9B to 36.07 against 38.35. The failure is specific to search-based QA, where one weak early query already moves the student off the assumed route. OPSD's teacher, conditioned on a fixed Gold plan, keeps generating continuations for a route the student never took, and every later step compounds the drift. SCOUT's gate instead reads the student's current state and, when the plan no longer fits, shifts weight toward the Skill and Neutral views, so the target tracks the student's actual position rather than a fixed reference. This state alignment contributes to SCOUT's gains across all benchmarks and backbones, while Gold-only OPSD performs worse than Vanilla on two Search-QA backbones. Appendix F shows two Qwen3.5-4B trajectories where this state-conditioned target resolves the specific failures.

Relative margins under GRPO differ by benchmark. Adding reward learning amplifies the advantage of SCOUT over every comparison that also uses GRPO. On ALFWorld, SCOUT+GRPO outperforms SDAR by 7.4, 6.2, and 0.5 points and GRPO by 9.0, 4.0, and 4.5 across the three backbones. On WebShop success, it outperforms OCSD by 6.5, 14.7, and 11.1 points and Gold-only OPSD+GRPO by 12.7, 8.7, and 12.8. On search-QA EM, it outperforms Gold-only OPSD+GRPO by 0.55, 4.19, and 0.87 points and OCSD by 1.03, 2.48, and 0.94. The pattern follows from what each method modifies. SDAR gates an auxiliary distillation loss with detached token-level signals, and OCSD modulates token-level updates by comparing replay views with and without future observations; neither changes the teacher target itself. SCOUT instead reweights the Gold, Skill, and Neutral teacher distributions based on the student's current state, so the target tracks the student's actual position. The margins are therefore largest on ALFWorld and WebShop, where trajectories can drift far from the assumed route, and narrowest on search-based QA, where the reward signal already corrects most of what the target would correct.

4.3 Training dynamics

Pure distillation. In Figure 4 (left), Gold-only OPSD and SCOUT stay within a narrow band through the first thirty steps, after which SCOUT pulls ahead and holds the lead through the final evaluations. Gold-only OPSD declines between steps 90 and 120 and recovers part of the distance afterwards, but never closes the gap. The timing matches the mechanism: early in training the stu-

Table 2: Component ablations with Qwen3.5-4B under pure distillation. Drops are relative to full SCOUT.

	ALFWorld success		Search-QA EM	
Variant	Result	Drop	Result	Drop
SCOUT	$\mathbf{48.1} \pm 0.8$	—	$\mathbf{42.28} \pm 0.02$	—
Fixed weights	45.9 ± 1.1	2.2	37.61 ± 0.13	4.67
w/o Gold view	41.7 ± 1.3	6.4	34.79 ± 0.10	7.49
w/o Skill view	37.6 ± 1.7	10.5	33.06 ± 0.19	9.22
w/o Neutral view	43.7 ± 1.2	4.4	37.93 ± 0.05	4.35

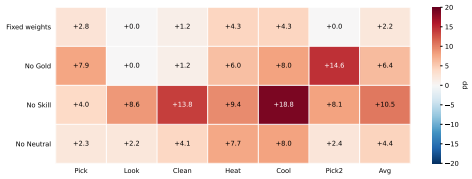

Figure 5: **Different views support different task requirements.** ALFWorld task-family success drops from full SCOUT to each Qwen3.5-4B pure-distillation ablation.

dent visits few states, so the gate has little reason to shift weight away from the Gold plan. Once the student reaches states that no longer match the assumed route, the gate shifts weight toward the Skill and Neutral views, and the advantage of state-conditioned weighting appears. **GRPO-augmented training.** The right panel shows the same pattern under reward learning: SCOUT+GRPO leads from step 30, Gold-only OPSD+GRPO briefly catches up near step 100 as reward covers more of the error surface, and SCOUT+GRPO separates again after step 110 as GRPO begins to exploit the better continuations supplied by the state-conditioned target. Appendix B.1 traces this late separation to the state-conditioned target, which keeps entropy from collapsing so exploration continues after OPSD+GRPO saturates.

4.4 Ablation Study

Adaptive weighting. Replacing the dynamic weights with a fixed 1:1:1 mixture lowers ALFWorld success by 2.2 points and search-QA EM by 4.67 points. The drop concentrates on Heat (4.3 points) and Cool (4.3 points), while Look and Pick2 are unchanged, showing that the three views are not interchangeable. These families require reaching a particular intermediate state before the final action, and a fixed mixture cannot shift weight toward Skill when the Gold plan no longer applies. SCOUT's gate performs this shift: when Gold compatibility drops, it moves probability mass toward Skill and Neutral, so the target reflects the student's actual position rather than a fixed average.

Three views. Each view supplies guidance that the others do not replace, and the task families where each one is removed show what it carries. The **Skill view** contributes most: removing it costs 10.5 points of ALFWorld success and 9.22 points of search-QA EM, concentrated on Cool (18.8), Clean (13.8), and Heat (9.4). Each of these families requires the object to reach a particular state before final placement, which is exactly what a skill stores in its preconditions and effects, so the student loses the guidance that completes intermediate transformations (Figure 5). Removing the **Gold view** costs 6.4 and 7.49 points, with the damage concentrated instead on Pick2 (14.6). Pick2 succeeds only when both objects reach the destination within one episode, so a policy that completes one valid subgoal can still fall short of the goal, and the global objective carried by Gold is what keeps the second object in sight. Removing the **Neutral view** costs 4.4 and 4.35 points spread across families, with the largest effects on Cool (8.0) and Heat (7.7). Neutral preserves action structure without instance-specific privileged content, matching its role as a structural fallback when the gate assigns little weight to Gold or Skill. The two privileged views therefore concentrate on different task families, and Neutral absorbs the states that neither covers. Appendix E separates the search-QA ablations into in-distribution and out-of-distribution groups.

5 Conclusion

Privileged supervision for multi-turn agents fails when a task-correct plan no longer applies to the state the student has reached. SCOUT solves this by adapting the teacher target to the student's state: a parameter-free gate reads compatibility and applicability signals and reweights Gold, Skill, and Neutral distributions before they reach the student. On ALFWorld, WebShop, and search-based QA, SCOUT improves over Gold-only OPSD on every backbone, both alone and when combined with GRPO. Removing any one view or fixing the gate weights removes the share of the gain, which confirms that state-conditioned composition is the mechanism that drives the improvement. Privileged supervision for multi-turn agents should follow the learner's state, not a fixed route.

ETHICS STATEMENT

We affirm our commitment to the ethical conduct of this research and provide the following assurances:

- This research was conducted in strict adherence to the highest ethical standards, and all findings have been reported with integrity, ensuring clarity and accuracy throughout our communications.
- Our study strictly avoids the use of sensitive or confidential data, ensuring that all materials are appropriate for public dissemination.
- The datasets employed in our experiments are sourced from publicly accessible and peer-reviewed scientific resources, ensuring transparency and reliability.
- We provided a detailed description of the dataset characteristics and the hyper-parameter settings used in our experiments to maintain transparency and consistency with our results.
- To promote transparency and facilitate further research, we commit to sharing our code in supplementary material and will open source after our paper is accepted.

REPRODUCIBILITY STATEMENT

Section 3 specifies the full SCOUT method, including the three teacher-conditioning contexts, the online skill buffer, the state signals, and the parameter-free gate. Appendices C and D provide the complete view definitions, buffer updates, signal extraction, and derivations. Training configurations for all three benchmarks appear in Appendix A, prompt templates in Appendix G, and per-task results behind every reported aggregate in Appendices B and E. Evaluation uses the student alone, with no privileged plan, skill retrieval, or teacher calls. Our complete training and evaluation code, with environment configurations and per-benchmark launch scripts, is provided as supplementary material.

AI USE STATEMENT

We used Codex as a writing assistant to help rewrite and polish the prose of the Introduction and Related Work sections for clarity and readability. The tool was not used to generate research ideas, design experiments, produce results, or draft the technical content of the Method, Experiments, or Analysis sections. All AI-suggested edits to the Introduction and Related Work were reviewed, fact-checked against our own sources, and revised by the authors, who take full responsibility for the accuracy, originality, and correctness of the final text.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5be69a584901a26c521c2b51e40a4c20-Abstract-Conference.html.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. PAL: Program-aided language models. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 10764–10799. PMLR, 2023. URL https://proceedings.mlr.press/v202/gao23f.html.

John Gkountouras, Josip Jukić, and Ivan Titov. Consensus as privileged context for label-free self-distillation. *arXiv preprint arXiv:2607.13643*, 2026. URL https://arxiv.org/abs/2607.13643.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, 2020. doi: 10.18653/v1/2020.coling-main.580. URL https://aclanthology.org/2020.coling-main.580/.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Noah Brown, Tomas Jackson, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. Inner monologue: Embodied reasoning through planning with language models. In *Proceedings of the 6th Conference on Robot Learning*, pp. 1769–1782. PMLR, 2023. URL https://proceedings.mlr.press/v205/huang23c.html.

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. *arXiv preprint arXiv:2503.09516*, 2025. URL https://arxiv.org/abs/2503.09516.

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

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, 2016. doi: 10.18653/v1/D16-1139. URL https://aclanthology.org/D16-1139/.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025. URL https://arxiv.org/abs/2503.20783.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015. URL https://arxiv.org/abs/1511.03643.

Zhengxi Lu, Zhiyuan Yao, Zhuowen Han, Zi-Han Wang, Jinyang Wu, Qi Gu, Xunliang Cai, Weiming Lu, Jun Xiao, Yueting Zhuang, and Yongliang Shen. Self-distilled agentic reinforcement learning. *arXiv preprint arXiv:2605.15155*, 2026. URL https://arxiv.org/abs/2605.15155.

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 *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, 2023. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546/.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 627–635. PMLR, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/d842425e4bf79ba039352da0f658a906-Abstract-Conference.html.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*, 2023. URL https://arxiv.org/abs/2303.11366.

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

Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. R1-Searcher: Incentivizing the search capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2503.05592*, 2025a. URL https://arxiv.org/abs/2503.05592.

Huatong Song, Jinhao Jiang, Wenqing Tian, Zhipeng Chen, Yuhuan Wu, Jiahao Zhao, Yingqian Min, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. R1-Searcher++: Incentivizing the dynamic knowledge acquisition of LLMs via reinforcement learning. *arXiv preprint arXiv:2505.17005*, 2025b. URL https://arxiv.org/abs/2505.17005.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31/.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=ehfRiF0R3a.

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. *arXiv preprint arXiv:2604.10674*, 2026. URL https://arxiv.org/abs/2604.10674.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

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

Yi Yang, Cong Qin, Xiaodan Liu, Chishui Chen, Qing Dong, Yan Zhang, Cao Liu, Zhao Yang, Lu Pan, Jiaye Lin, and Yi Feng. Agentic reinforcement learning with observation-calibrated self-distillation. *arXiv preprint arXiv:2608.04788*, 2026b. URL https://arxiv.org/abs/2608.04788.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. WebShop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2207.01206.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

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 *Advances in Neural Information Processing Systems*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/a4277440d50f1f15d2cb4c14f7e0c0d2-Abstract-Conference.html.

Guibin Zhang, Hejia Geng, Xiaohang Yu, Zhenfei Yin, Zaibin Zhang, Zelin Tan, Heng Zhou, Zhong-Zhi Li, Xiangyuan Xue, Yijiang Li, Yifan Zhou, Yang Chen, Chen Zhang, Yutao Fan, Zihu Wang, Songtao Huang, Francisco Piedrahita Velez, Yue Liao, Hongru Wang, Mengyue Yang, Heng Ji, Jun Wang, Shuicheng Yan, Philip Torr, and Lei Bai. The landscape of agentic reinforcement learning for LLMs: A survey. *Transactions on Machine Learning Research*, 2026a. URL https://openreview.net/forum?id=RY19y2RI1O.

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

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. ExpeL: LLM agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19632–19642, 2024. doi: 10.1609/aaai.v38i17.29936. URL https://ojs.aaai.org/index.php/AAAI/article/view/29936.

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. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Jpxfof0EaS.

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*, 2026a. URL https://arxiv.org/abs/2606.19659.

Yuhang Zhou, Kai Zheng, Haoling Li, Dengyun Peng, Can Xu, and Jingjing Chen. TurnOPD: Making on-policy distillation turn-aware for efficient long-horizon agent training. *arXiv preprint arXiv:2607.05804*, 2026b. URL https://arxiv.org/abs/2607.05804.

A Training Details

All three benchmarks share the same optimizer and core schedule. We use AdamW with a learning rate of 1×10^{-6} and train for 150 steps on every benchmark. The per-task differences reflect environment constraints rather than per-task tuning. ALFWorld and WebShop use a task batch size of 16, while Search-QA uses 128 because each episode is a short single-query interaction. Maximum interaction steps follow each environment's horizon: 50 for ALFWorld, 15 for WebShop, and 4 for Search-QA. Prompt length is capped at 2048 tokens on ALFWorld and 4096 on WebShop and Search-QA, matching the observation and retrieval lengths of each task and are listed in Table 3.

Table 3: Training hyperparameters for each benchmark.

Hyperparameter	**ALFWorld**	**WebShop**	**Search-QA**
Task batch size	16	16	128
Learning rate	1×10^{-6}	1×10^{-6}	1×10^{-6}
Maximum interaction steps	50	15	4
Maximum prompt length	2048	4096	4096
Training steps	150	150	150
Evaluation numbers	140 (seen) + 134 (unseen)	500	51,713

B Detailed Evaluation Results

The following tables retain the full reported precision and evaluation-seed variance underlying Table 1. Values without variance are point estimates. OPSD denotes Gold-only OPSD. ID and OOD follow the dataset groups in the experimental record. NQ, Triv, Pop, Hotp, 2Wk, MuS, and Bam abbreviate Natural Questions, TriviaQA, PopQA, HotpotQA, 2WikiMultihopQA, MuSiQue, and Bamboogle, respectively. A short analysis follows each table.

Table 4 gives the full breakdown for Qwen3-4B. SCOUT+GRPO leads every aggregate column, reaching 63.0 on ALFWorld, 47.30 on Search-QA, and 87.0 / 75.5 on WebShop score and success. The pure-distillation comparison already favors SCOUT on all three benchmarks, with the largest relative gain on WebShop success, 28.2 against 8.8 for Gold-only OPSD. On Search-QA the dataset-level rows show the gain is broad rather than concentrated: SCOUT improves over OPSD on all seven datasets, with the largest margins on TriviaQA (60.33 vs. 51.35) and HotpotQA (33.96 vs. 26.09) and the only near-tie on PopQA (36.00 vs. 35.68). The gains cover single-hop and multi-hop datasets alike, so state-conditioned recovery guidance improves retrieval-augmented reasoning across question types rather than on one structural family.

Qwen3.5-4B (Table 5) widens every gap seen on Qwen3-4B. Pure SCOUT raises the ALFWorld average from 44.8 to 48.1 and the WebShop score from 30.3 to 51.1 over Gold-only OPSD, and on Search-QA it turns OPSD's below-Vanilla regression (31.00 vs. 33.57) into a 42.28 aggregate. The GRPO-augmented setting makes the separation decisive: SCOUT+GRPO reaches 93.2 on ALFWorld and 49.93 on Search-QA, ahead of OPSD+GRPO by 5.0 and 4.19 points and ahead of SDAR, the strongest loss-reweighting baseline, by 6.2 and 4.30 points. The dataset-level rows show where the advantage concentrates. On 2WikiMultihopQA, SCOUT+GRPO reaches 47.77 against 39.56 for OPSD+GRPO, the largest per-dataset margin in this block, consistent with the Skill view supplying the query-reformulation structure that multi-hop questions reward.

Qwen3.5-9B (Table 6) tests whether the method survives a stronger base policy. It does. SCOUT+GRPO reaches 93.9 on ALFWorld, 51.34 on Search-QA, and 89.1 / 77.8 on WebShop score and success, the highest aggregate in every column of this block. The pure-distillation comparison is again the sharpest on Search-QA, where SCOUT improves over Gold-only OPSD by 6.97 points while OPSD itself falls below Vanilla. The family-level rows also show a backbone interaction worth noting: on Heat, the 9B backbone rises from 24.8 under pure distillation to 83.8 with GRPO, the largest single-cell gain in the table, because the stronger base policy converts the Skill view's precondition guidance into successful heating trajectories far more reliably once reward learning is active.

Table 4: Detailed evaluation results for Qwen3-4B. Values are mean ± evaluation-seed variance where available.

ALFWorld success and WebShop results

Method	Pick	Look	Clean	Heat	Cool	Pick2	Avg	WS score	WS succ.
Vanilla	69.5	32.3	22.4	17.9	8.7	24.4	31.0	14.5	4.8
OPSD	79.7 ± 1.7	24.7 ± 3.7	42.5 ± 3.6	28.2 ± 5.1	15.2 ± 2.2	40.7 ± 5.1	41.6 ± 1.0	35.4 ± 0.41	8.8 ± 1.2
GRPO	89.3 ± 2.6	51.6 ± 3.2	32.8 ± 4.6	43.6 ± 6.8	30.4 ± 2.2	71.5 ± 3.7	54.0 ± 3.1	83.1 ± 0.4	66.0 ± 1.1
OPSD+GRPO	88.1 ± 1.7	53.8 ± 1.9	31.6 ± 1.0	53.0 ± 7.8	26.8 ± 1.3	59.3 ± 5.6	52.7 ± 2.0	81.6 ± 0.21	62.8 ± 0.60
OCSD	89.8 ± 1.7	51.6 ± 3.2	33.3 ± 1.0	53.0 ± 1.5	28.3 ± 2.2	59.3 ± 2.8	53.4 ± 1.1	83.7 ± 0.3	69.0 ± 0.2
SDAR	88.7 ± 2.0	37.6 ± 6.7	42.0 ± 4.3	55.6 ± 5.3	37.0 ± 2.2	61.8 ± 5.1	55.6 ± 1.1	79.2 ± 1.5	58.1 ± 2.7
SCOUT	79.1 ± 1.0	32.3 ± 3.2	41.4 ± 1.7	39.3 ± 5.3	15.9 ± 4.5	40.7 ± 1.4	43.8 ± 1.5	40.8 ± 0.40	28.2 ± 0.35
SCOUT+GRPO	94.4 ± 1.0	74.2 ± 3.2	33.9 ± 1.0	68.4 ± 3.9	43.5 ± 4.3	67.5 ± 1.4	63.0 ± 0.4	87.0 ± 0.63	75.5 ± 0.76

Search-QA: dataset-level exact match

Method	NQ	Triv	Pop	Hotp	2Wk	MuS	Bam
Vanilla	29.78	49.58	30.90	26.89	30.63	7.20	25.60
OPSD	34.42 ± 0.44	51.35 ± 0.21	35.68 ± 0.46	26.09 ± 0.25	31.49 ± 0.19	7.76 ± 0.30	25.87 ± 1.49
GRPO	44.02 ± 0.31	63.91 ± 0.12	45.68 ± 0.08	41.66 ± 0.32	39.54 ± 0.03	17.17 ± 0.25	71.20 ± 0.64
OPSD+GRPO	43.85 ± 0.33	64.24 ± 0.04	46.69 ± 0.24	42.97 ± 0.09	39.58 ± 0.18	17.25 ± 0.41	72.00 ± 1.92
OCSD	43.82 ± 0.17	61.88 ± 0.30	47.16 ± 0.15	40.53 ± 0.10	40.90 ± 0.25	15.89 ± 0.25	71.20 ± 4.48
SDAR	41.02 ± 0.12	58.43 ± 0.14	45.69 ± 0.22	35.34 ± 0.38	38.55 ± 0.40	16.55 ± 0.23	68.80 ± 1.92
SCOUT	39.03 ± 0.77	60.33 ± 0.26	36.00 ± 0.23	33.96 ± 0.51	35.32 ± 0.21	12.16 ± 0.25	31.20 ± 4.48
SCOUT+GRPO	45.43 ± 0.37	64.56 ± 0.10	47.00 ± 0.31	42.62 ± 0.27	41.00 ± 0.36	17.00 ± 0.10	70.40 ± 2.56

Method	ID	OOD	Avg
Vanilla	27.83	34.59	33.15
OPSD	28.82 ± 0.23	37.05 ± 0.06	35.30 ± 0.01
GRPO	42.43 ± 0.24	47.23 ± 0.04	46.21 ± 0.05
OPSD+GRPO	43.26 ± 0.09	47.70 ± 0.07	46.75 ± 0.04
OCSD	41.61 ± 0.02	47.54 ± 0.10	46.27 ± 0.08
SDAR	37.20 ± 0.23	45.37 ± 0.20	43.63 ± 0.19
SCOUT	35.62 ± 0.15	41.12 ± 0.11	39.95 ± 0.08
SCOUT+GRPO	43.54 ± 0.20	48.32 ± 0.04	47.30 ± 0.07

Table 5: Detailed evaluation results for Qwen3.5-4B. Values are mean ± evaluation-seed variance where available.

ALFWorld success and WebShop results

Method	Pick	Look	Clean	Heat	Cool	Pick2	Avg	WS score	WS succ.
Vanilla	79.7	6.5	15.5	5.1	8.7	39.0	29.2	14.6	8.6
OPSD	83.6 ± 2.0	18.3 ± 4.9	41.4 ± 3.0	17.1 ± 1.5	23.2 ± 2.5	64.2 ± 1.4	44.8 ± 0.8	30.3 ± 1.3	19.3 ± 1.8
GRPO	93.8 ± 1.0	96.8 ± 3.2	86.8 ± 3.6	79.5 ± 2.6	87.7 ± 2.5	91.1 ± 2.8	89.2 ± 1.2	82.3 ± 0.5	64.0 ± 0.4
OPSD+GRPO	95.5 ± 1.0	76.3 ± 1.9	88.5 ± 2.0	82.9 ± 3.0	87.7 ± 1.3	91.9 ± 1.4	88.2 ± 1.1	84.5 ± 0.34	68.7 ± 0.12
OCSD	92.7 ± 4.3	93.5 ± 3.2	89.7 ± 1.7	65.8 ± 3.0	86.2 ± 1.3	82.9 ± 2.4	85.8 ± 0.4	81.9 ± 1.2	62.7 ± 0.23
SDAR	94.4 ± 2.6	73.1 ± 4.9	93.1 ± 1.7	77.8 ± 3.9	92.0 ± 1.3	81.3 ± 3.7	87.0 ± 1.5	80.1 ± 1.5	65.4 ± 1.1
SCOUT	85.9 ± 2.0	18.3 ± 1.9	44.3 ± 2.0	23.1 ± 2.6	32.6 ± 2.2	62.6 ± 1.4	48.1 ± 0.8	51.1 ± 1.5	33.8 ± 1.7
SCOUT+GRPO	96.0 ± 2.0	97.8 ± 3.7	92.5 ± 2.0	87.2 ± 2.6	93.5 ± 2.2	91.9 ± 1.4	93.2 ± 0.4	88.9 ± 0.26	77.4 ± 0.4

Search-QA: dataset-level exact match

Method	NQ	Triv	Pop	Hotp	2Wk	MuS	Bam
Vanilla	30.75	50.11	32.03	25.87	30.82	7.07	38.40
OPSD	24.40 ± 0.35	50.46 ± 0.26	24.34 ± 0.14	25.17 ± 0.19	31.08 ± 0.40	6.43 ± 0.40	33.33 ± 6.61
GRPO	45.40 ± 0.34	63.79 ± 0.16	44.03 ± 0.04	40.66 ± 0.08	40.78 ± 0.06	15.93 ± 0.21	70.40 ± 0.64
OPSD+GRPO	45.29 ± 0.38	60.37 ± 0.20	45.26 ± 0.33	44.00 ± 0.31	39.56 ± 0.49	16.88 ± 0.71	72.00 ± 0.64
OCSD	44.13 ± 0.40	62.99 ± 0.09	48.24 ± 0.14	42.55 ± 0.19	42.71 ± 0.24	13.74 ± 0.75	66.40 ± 2.56
SDAR	42.85 ± 0.54	60.29 ± 0.25	46.35 ± 0.11	40.58 ± 0.22	40.47 ± 0.18	18.45 ± 0.13	60.00 ± 0.64
SCOUT	40.03 ± 0.35	59.16 ± 0.15	38.87 ± 0.25	38.88 ± 0.41	39.27 ± 0.09	11.87 ± 0.15	60.80 ± 5.76
SCOUT+GRPO	46.51 ± 0.58	61.20 ± 0.33	50.50 ± 0.21	46.32 ± 0.28	47.77 ± 0.14	20.11 ± 1.09	71.20 ± 2.56

Method	ID	OOD	Avg
Vanilla	27.47	35.22	33.57
OPSD	24.92 ± 0.18	32.65 ± 0.11	31.00 ± 0.05
GRPO	42.22 ± 0.13	46.93 ± 0.04	45.93 ± 0.02
OPSD+GRPO	44.42 ± 0.25	46.10 ± 0.31	45.74 ± 0.28
OCSD	43.07 ± 0.15	48.64 ± 0.18	47.45 ± 0.17
SDAR	41.33 ± 0.11	46.80 ± 0.15	45.63 ± 0.09
SCOUT	39.26 ± 0.19	43.10 ± 0.04	42.28 ± 0.02
SCOUT+GRPO	46.38 ± 0.09	50.89 ± 0.18	49.93 ± 0.14

Table 6: Detailed evaluation results for Qwen3.5-9B. Values are mean ± evaluation-seed variance where available.

ALFWorld success and WebShop results

Method	Pick	Look	Clean	Heat	Cool	Pick2	Avg	WS score	WS succ.
Vanilla	78.0	54.8	20.7	25.6	4.3	48.8	39.1	22.9	15.2
OPSD	89.3 ± 3.5	49.5 ± 4.9	33.9 ± 2.0	60.7 ± 6.5	29.7 ± 2.5	65.0 ± 3.7	55.4 ± 1.3	40.8 ± 2.5	28.0 ± 2.0
GRPO	97.2 ± 1.0	95.7 ± 3.7	89.7 ± 3.4	76.9 ± 6.8	81.2 ± 2.5	94.3 ± 5.6	89.4 ± 1.1	80.2 ± 1.8	65.5 ± 1.1
OPSD+GRPO	93.2 ± 1.7	93.5 ± 3.2	83.3 ± 1.0	77.8 ± 1.5	84.8 ± 2.2	94.3 ± 1.4	87.7 ± 0.4	78.3 ± 0.2	65.0 ± 1.0
OCSD	92.1 ± 2.6	92.5 ± 3.7	89.1 ± 5.0	71.8 ± 6.8	82.6 ± 5.8	84.6 ± 2.8	85.9 ± 0.4	82.2 ± 0.19	66.7 ± 2.0
SDAR	96.0 ± 2.0	79.6 ± 6.7	97.7 ± 1.0	88.9 ± 1.5	96.4 ± 3.3	95.1 ± 2.4	93.4 ± 1.0	83.1 ± 0.2	67.1 ± 0.6
SCOUT	90.4 ± 2.0	65.6 ± 1.9	65.5 ± 1.7	24.8 ± 3.0	34.1 ± 3.3	69.9 ± 5.1	60.5 ± 0.4	55.5 ± 1.3	38.2 ± 1.2
SCOUT+GRPO	96.6 ± 2.9	98.9 ± 1.9	98.3 ± 1.7	83.8 ± 1.5	92.0 ± 2.5	91.9 ± 1.4	93.9 ± 0.6	89.1 ± 0.7	77.8 ± 1.9

Search-QA: dataset-level exact match

Method	NQ	Triv	Pop	Hotp	2Wk	MuS	Bam
Vanilla	30.97	54.85	33.73	46.32	31.83	8.77	40.00
OPSD	29.70 ± 0.33	56.36 ± 0.44	31.19 ± 0.12	29.43 ± 0.09	33.77 ± 0.25	11.54 ± 0.50	40.80 ± 4.48
GRPO	45.98 ± 0.56	62.37 ± 0.35	47.54 ± 0.11	41.61 ± 0.41	40.31 ± 0.09	17.38 ± 0.25	71.20 ± 2.56
OPSD+GRPO	46.84 ± 0.29	65.34 ± 0.13	50.71 ± 0.08	46.55 ± 0.21	45.88 ± 0.13	19.61 ± 0.46	72.00 ± 2.56
OCSD	47.89 ± 0.37	66.99 ± 0.04	49.80 ± 0.35	45.12 ± 0.39	45.60 ± 0.07	20.02 ± 0.15	71.20 ± 0.64
SDAR	48.06 ± 0.48	61.61 ± 0.31	47.70 ± 0.08	47.00 ± 0.42	40.36 ± 0.22	18.99 ± 0.59	72.80 ± 1.92
SCOUT	45.37 ± 0.48	62.87 ± 0.30	37.05 ± 0.11	39.32 ± 0.43	38.07 ± 0.40	18.78 ± 0.33	52.80 ± 2.56
SCOUT+GRPO	48.17 ± 0.62	65.55 ± 0.18	51.80 ± 0.17	47.72 ± 0.30	46.66 ± 0.29	21.10 ± 0.05	72.80 ± 1.92

Method	ID	OOD	Avg
Vanilla	41.29	37.55	38.35
OPSD	29.51 ± 0.15	37.85 ± 0.08	36.07 ± 0.10
GRPO	43.04 ± 0.30	47.71 ± 0.12	46.72 ± 0.15
OPSD+GRPO	46.65 ± 0.19	51.50 ± 0.08	50.47 ± 0.05
OCSD	46.03 ± 0.36	51.58 ± 0.11	50.40 ± 0.07
SDAR	47.34 ± 0.19	47.67 ± 0.10	47.60 ± 0.10
SCOUT	41.31 ± 0.14	43.51 ± 0.15	43.04 ± 0.13
SCOUT+GRPO	47.87 ± 0.26	52.28 ± 0.06	51.34 ± 0.10

Across the three backbones, the pattern in Table 1 holds at full precision: pure SCOUT exceeds Gold-only OPSD on the aggregate for ALFWorld, Search-QA, and WebShop for Qwen3-4B, Qwen3.5-4B, and Qwen3.5-9B alike, and SCOUT+GRPO exceeds OPSD+GRPO by the same pattern. The comparison is not only a matter of degree: on Search-QA, Gold-only OPSD leaves the aggregate exact match below the untrained Vanilla backbone for Qwen3.5-4B (31.00 vs. 33.57) and Qwen3.5-9B (36.07 vs. 38.35), even though the same target still raises the ALFWorld and WebShop averages. Search-QA trajectories involve several sequential queries before an answer is accepted, so an early suboptimal query already leaves the student outside the single successful search path that Gold assumes; matching a token distribution conditioned on that path can then pull the student toward actions that do not fit its own history. SCOUT+GRPO also reaches the highest Search-QA and ALFWorld aggregate among OPSD+GRPO, OCSD, and SDAR for all three backbones.

B.1 Response Length and Entropy Dynamics

Figure 6 tracks mean response length and policy entropy for Qwen3.5-9B under GRPO-augmented training. The two panels expose a compound advantage. On the length axis, SCOUT+GRPO converges to a lower and still-declining response length than Gold-only OPSD+GRPO, as the state-conditioned target teaches the student to reach the task goal with fewer tokens and GRPO continues to compress the response as reward accumulates. On the entropy axis, the same target prevents the entropy collapse that afflicts Gold-only OPSD+GRPO. Under Gold-only OPSD+GRPO, entropy decreases monotonically toward zero as the student is pulled toward a fixed plan and sharpens onto a narrowing set of continuations until the policy becomes nearly deterministic. Under SCOUT+GRPO, entropy contracts early but then recovers and stays at a higher level, because the state-conditioned gate keeps assigning mass to applicable Skill and Neutral continuations when the student's state departs from the assumed route. A policy that retains entropy keeps the exploration capacity that group-relative updates need to keep improving late in training, and Figure 4 shows the corresponding score separation over the same period. The two panels together show that the gate's state-dependent weighting produces both efficiency and robustness, yielding shorter responses without the entropy collapse that limits Gold-only OPSD+GRPO.

C Method Details

C.1 View Definitions

Gold view. The context $z_t^G = z^G(x)$ contains a successful plan for the task. It supplies the global task route, but need not describe every recovery step after a student mistake. Its token distribution still changes with h_t and $y_{t,<l}$, even though the privileged plan remains fixed.

Skill view. The context z_t^S contains a reusable skill retrieved from the training-time buffer. A skill specifies preconditions, an action schema, expected effects, and recovery or re-entry rules. These rules describe how to resume an unfinished subgoal from a compatible intermediate state, rather than requiring the student to reproduce a complete demonstration from its beginning.

Neutral view. The context z_t^N retains the action structure of the privileged plan while removing instance-specific entities and solution content. The original task and student history remain available through h_t. Neutral therefore removes a source of privileged guidance. It retains the action structure and the student history, and it provides a structural fallback when the other views receive little weight.

C.2 Skill Buffer Construction and Update

A cleaning skill, for example, can require the target object to be held before cleaning, and include a re-entry rule for acquiring the correct object after a mistaken pickup. These are teacher-conditioning instructions, not hard constraints on the student's actions.

The buffer starts from historical successful training trajectories and grows from new successes. A frozen teacher abstracts environment-confirmed successful interactions into skill descriptions, which are validated and stripped of instance-specific identifiers before insertion. If $\mathcal{T}_k$ is the current rollout

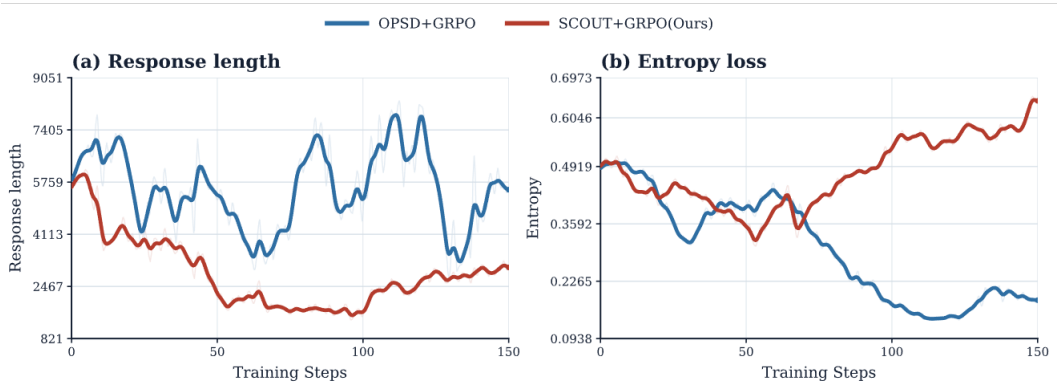

Figure 6: **Response length and entropy loss during training.** (a) Response length decreases for both methods, with SCOUT+GRPO consistently producing shorter responses. (b) Entropy keeps decreasing under OPSD+GRPO, while SCOUT+GRPO maintains a higher, rising entropy, mitigating entropy collapse.

batch, the update is

$$\begin{aligned} \mathcal{K}_k &= \{\mathrm{Extract}_\phi(\tau) : \tau \in \mathcal{T}_k,\ \mathrm{Success}(\tau) = 1\}, \\ \mathcal{B}_{k+1} &= \mathrm{Update}(\mathcal{B}_k, \mathrm{Validate}(\mathcal{K}_k)). \end{aligned} \tag{10}$$

The snapshot $\mathcal{B}_k$ remains fixed while the current batch is scored. Newly accepted skills become available only in the next iteration, so a trajectory cannot supply its own newly extracted skill as supervision in the same update. Evaluation trajectories never update the buffer.

C.3 State-Signal Extraction

The endpoint-preserving sigmoid map g_{T_g} smooths compatibility with temperature T_g, while satisfying $g_{T_g}(0) = 0$ and $g_{T_g}(1) = 1$. In ALFWorld, the signals use object identity and task-precondition information, including inventory and transformation status. The semantic roles of the signals and the mixing rule are shared across environments, but their extraction rules are not learned or assumed to transfer unchanged. If no skill is retrieved, we set its applicability to zero.

C.4 Numerical Evaluation of the Mixture

For the sampled token, the mixture is evaluated stably in log space:

$$\log q^\star_{t,l}(y_{t,l}) = \log \sum_{v:\, \bar{w}^v_{t,l} > 0} \exp\big(\log \bar{w}^v_{t,l} + \log q^v_{t,l}(y_{t,l})\big)\,. \tag{11}$$

This computes the exact sampled-token probability of the arithmetic mixture in Eq. 7. Caching three scalar log-probabilities at each of N scored positions requires $O(3N)$ storage instead of caching three full $|\mathcal{Y}|$-dimensional vocabulary distributions at those positions. This does not remove the teacher forward passes or their temporary vocabulary computations.

D Properties of the Gate and Mixture

D.1 Gate Normalization and Boundary Cases

Normalization and fallback. The Neutral weight has the equivalent decomposition

$$w^N_t = \eta(1 - \alpha_t) + b_t(1 - \alpha_t)(1 - \beta_t). \tag{12}$$

The first term reserves Neutral weight as Gold compatibility decreases; the second receives mass that is supported by neither Gold compatibility nor skill applicability. All factors are nonnegative,

and the three weights sum to one. The coefficient η is therefore not a fixed Neutral floor at every state: the reserve vanishes when $\alpha_t = 1$.

The gate has three useful boundary cases:

$$(w_t^G, w_t^S, w_t^N) = \begin{cases} (1,0,0), & \alpha_t = 1, \\ (0, 1-\eta, \eta), & \alpha_t = 0,\ \beta_t = 1, \\ (0,0,1), & \alpha_t = 0,\ \beta_t = 0. \end{cases} \tag{13}$$

The first recovers the Gold-only teacher target. The second retains applicable skill guidance when Gold is incompatible, while the third uses Neutral alone. Moreover, $\partial w_t^G / \partial \alpha_t = 1 - \eta + 2\eta\alpha_t \geq 0$ and $\partial w_t^S / \partial \beta_t = b_t(1 - \alpha_t) \geq 0$. Thus, increasing either signal cannot decrease its corresponding view's weight when the other quantities are fixed.

D.2 Mixing Distributions versus Mixing Losses

Why mix before distilling? For any fixed prefix, student distribution p, normalized weights $\bar{w}_v$, and positive teacher distributions, convexity of $-\log$ gives

$$D_{\mathrm{KL}}\left(p \,\middle\|\, \sum_v \bar{w}_v q_v\right) \leq \sum_v \bar{w}_v D_{\mathrm{KL}}(p \| q_v). \tag{14}$$

The inequality follows by applying $-\log \sum_v \bar{w}_v q_v(u) \leq -\sum_v \bar{w}_v \log q_v(u)$ at each token and averaging under p. It explains why a mixture target differs from requiring agreement with each view through separate weighted losses: one view can support a token that another assigns low probability.

D.3 Sampled Distillation Surrogate

Let $\mathcal{C}_k$ denote the empirical distribution of loss-bearing prefixes collected by the rollout policy π_{θ_k}. The distribution-level reference objective is

$$J_{\mathrm{dist}}(\theta; \mathcal{C}_k) = \mathbb{E}_{c \sim \mathcal{C}_k}\left[D_{\mathrm{KL}}\left(\pi_\theta(\cdot \mid c) \,\|\, \mathrm{sg}[q^\star(\cdot \mid c)]\right)\right], \tag{15}$$

where sg stops gradients through the mixed teacher target. Prefixes in $\mathcal{C}_k$ are held fixed during each optimizer update.

For a sampled token $y_{t,l} \sim p_{\theta_k,t,l}$, we use the low-variance surrogate

$$d_{t,l} = \log \mathrm{sg}\left[q^\star_{t,l}(y_{t,l})\right] - \log p_{\theta,t,l}(y_{t,l}), \qquad \ell_{t,l} = \exp(d_{t,l}) - 1 - d_{t,l}, \tag{16}$$

and compute

$$L_{\mathrm{dist}} = \mathrm{Reduce}_m\left(\{\ell_{t,l}\}_{t,l}\right), \tag{17}$$

where Reduce_m is the baseline masked aggregation over loss-bearing tokens. Numerical clipping and normalization are identical for SCOUT and the Gold-only comparator; only the teacher probability is replaced by $q^\star$.

For normalized distributions p and q with shared support,

$$\mathbb{E}_{y \sim p}\left[\frac{q(y)}{p(y)} - 1 + \log \frac{p(y)}{q(y)}\right] = D_{\mathrm{KL}}(p \| q). \tag{18}$$

Thus, Eq. 16 provides a sampled surrogate for the reverse-KL reference objective in Eq. 15 under current-policy sampling (Wang et al., 2026).

E Additional Ablation Results

Table 7 provides the means and evaluation-seed variance underlying Table 2 and Figure 5. Values are mean $\pm$ evaluation-seed variance. ALFWorld reports success percentages, and Search-QA reports exact-match percentages.

Table 7: Qwen3.5-4B pure-distillation ablations. Values are mean $\pm$ evaluation-seed variance. ALFWorld reports success percentages. Search-QA reports exact-match percentages.

	Pick	Look	Clean	Heat	Cool	Pick2	Avg
Full	85.9 ± 2.0	18.3 ± 1.9	44.3 ± 2.0	23.1 ± 2.6	32.6 ± 2.2	62.6 ± 1.4	48.1 ± 0.8
Fixed weights	83.1 ± 1.7	18.3 ± 1.9	43.1 ± 1.7	18.8 ± 3.0	28.3 ± 2.2	62.6 ± 1.4	45.9 ± 1.1
w/o Gold	78.0 ± 2.9	18.3 ± 3.7	43.1 ± 1.7	17.1 ± 3.0	24.6 ± 2.5	48.0 ± 2.8	41.7 ± 1.3
w/o Skill	81.9 ± 2.0	9.7 ± 3.2	30.5 ± 2.6	13.7 ± 1.5	13.8 ± 3.3	54.5 ± 3.7	37.6 ± 1.7
w/o Neutral	83.6 ± 2.0	16.1 ± 3.2	40.2 ± 2.0	15.4 ± 2.6	24.6 ± 1.3	60.2 ± 2.8	43.7 ± 1.2

	QA ID	QA OOD	QA Avg
Full	39.26 ± 0.19	43.10 ± 0.04	42.28 ± 0.02
Fixed weights	32.47 ± 0.14	39.00 ± 0.15	37.61 ± 0.13
w/o Gold	29.06 ± 0.49	36.34 ± 0.06	34.79 ± 0.10
w/o Skill	28.92 ± 0.61	34.18 ± 0.14	33.06 ± 0.19
w/o Neutral	33.32 ± 0.19	39.17 ± 0.08	37.93 ± 0.05

Adaptive weighting across task families. Fixed 1:1:1 weights mainly lower Heat and Cool success by 4.3 and 4.3 points, while leaving Look and Pick2 unchanged. These task families require the student to reach a suitable intermediate state before the final action. A fixed mixture cannot shift weight toward the Skill view when the Gold plan no longer fits the reached state. The state-dependent gate makes this shift through Gold compatibility and skill applicability.

View-specific ALFWorld patterns. Removing the Skill view causes the largest drops on Cool (18.8 points), Clean (13.8 points), and Heat (9.4 points), but lowers Pick by only 4.0 points. Cool, Clean, and Heat require intermediate object-state changes before the final placement step, matching the preconditions and effects encoded in the Skill view. Pick is determined mainly by the global objective, so it is less affected by removing the Skill view. In contrast, removing the Gold view is most costly on Pick2 (14.6 points). Pick2 requires completing two object-placement subgoals within one episode, so the Gold view helps retain the second subgoal after the first is completed. The Gold and Skill views therefore have distinct task-family profiles. Removing the Neutral view produces smaller losses across all task families, largest on Cool (8.0 points) and Heat (7.7 points). This broad pattern is consistent with Neutral serving as a structural fallback rather than supporting one task family alone.

Search-QA ID/OOD patterns. Table 7 also separates the Search-QA ablation into in-distribution and out-of-distribution groups, a split not shown in the main text. Removing Skill produces the largest drop on both groups (10.34 points ID, 8.92 points OOD), while removing Gold produces a comparably large ID drop (10.20 points) but a smaller OOD drop (6.76 points). If Gold's task plan were the main driver of generalization to unseen question types, its removal would be expected to cost more OOD than Skill's removal. The opposite pattern holds here. This is consistent with the Skill view supplying reusable query-reformulation and evidence-gathering structure that transfers across datasets, rather than a plan tied to one dataset's question format. Keeping all three views but replacing the parameter-free gate with fixed 1:1:1 weights still costs 4.67 points on the Search-QA aggregate, and the loss is larger in distribution (6.79 points) than out of distribution (4.10 points). Access to the three views is therefore not sufficient on its own: the state-dependent weighting that decides when each view applies is itself a source of the gain.

F CASE STUDY

These Qwen3.5-4B examples were selected from tasks where SCOUT succeeds and Gold-only OPSD fails. The publication copies retain the observations and actions from the recorded trajectories. Student-only evaluation has no active gate or skill retrieval.

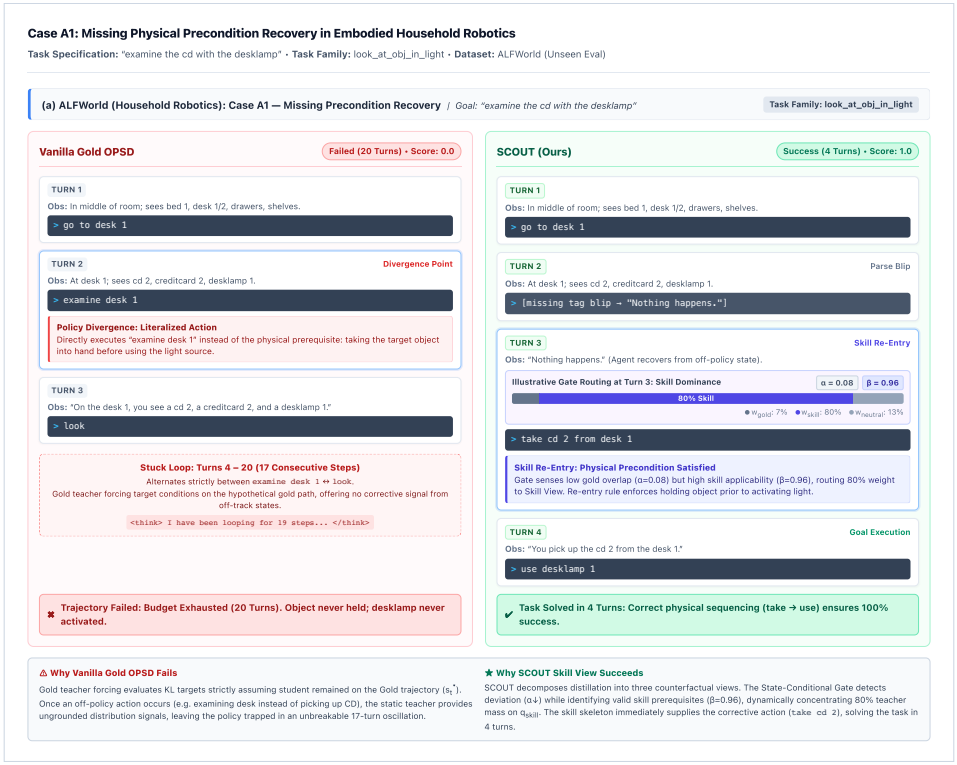

Figure 7: **Case A1.** Gold alternates examination and looking without acquiring the CD. SCOUT takes the CD and uses the lamp after a malformed response.



Figure 8: **Case W1.** Both reach the same product and select attributes. Gold later resets that state through navigation. SCOUT purchases after selection.

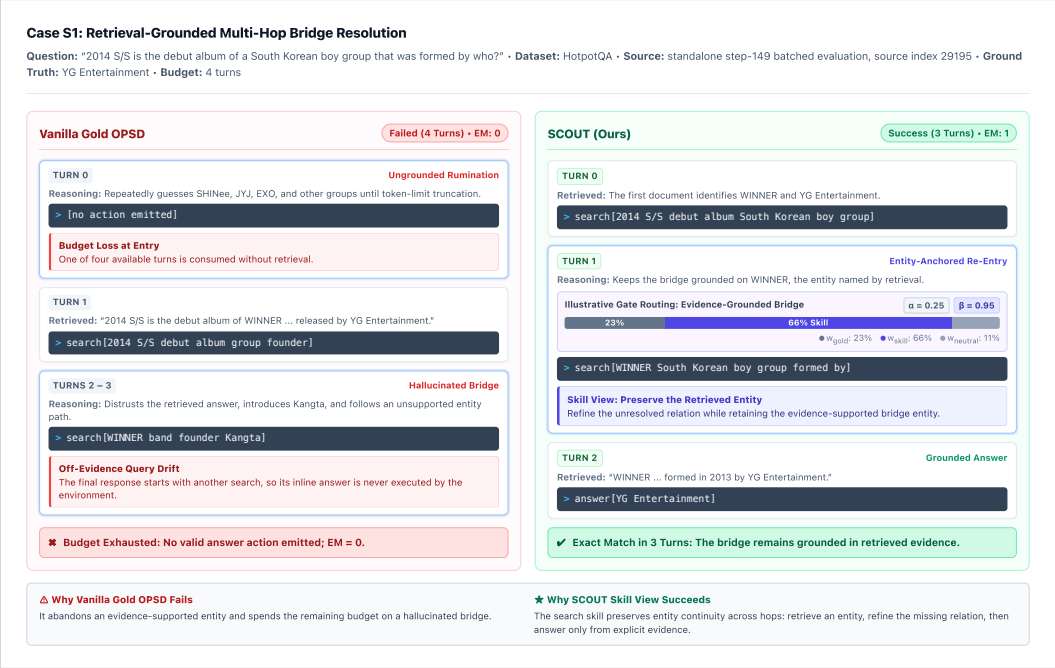

Figure 9: **Case S1.** The policies differ in whether they retain and use the entity named in retrieved evidence.

G PROMPT TEMPLATE

ALFWORLD_TEMPLATE

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

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

Figure 10: Prompt of policy model in ALFWorld.

WEBSHOP_TEMPLATE

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

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

Figure 11: Prompt of policy model in Webshop.

SEARCH_TEMPLATE

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

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

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

Figure 12: Prompt of policy model in Search-based QA.