Trust Is Not Enough: Influence Calibration for On-Policy Self-Distillation in Agentic RL

Qizhen Lan[1]*, Xi Xiao[2]*, Xiangchen Guan[3]*, Mengchen Fan[2], Moule Lin[4], Jung Im Choi[5], Lijing Zhu[6]†

[1]The University of Texas Health Science Center at Houston
[2]University of Alabama at Birmingham
[3]Amazon
[4]Trinity College Dublin and Lero
[5]The University of Findlay
[6]University of Houston-Clear Lake
Qizhen.Lan@uth.tmc.edu, Zhul@uhcl.edu

Abstract

On-policy self-distillation (OPSD) gives language agents dense token-level supervision from a privileged self-teacher on the policy's own trajectories. Existing methods allocate this supervision mainly by teacher trust, but trust does not reveal whether emphasizing a token supports the current policy objective. We call this the trust–utility mismatch and introduce Influence Calibration for Self-Distillation (ICSD). For each supervised token, ICSD measures the first-order response of its importance-weighted RL surrogate contribution to a teacher-directed output perturbation. Batch-adaptive calibration converts this non-stationary signal into a bounded allocation weight while preserving the original auxiliary-loss mass within each action turn. These detached weights affect only the distillation loss and require no additional model pass. Across ALFWorld, WebShop, and Search-QA, ICSD improves all matched aggregate metrics over trust-only allocation under Group Relative Policy Optimization (GRPO) and Group-in-Group Policy Optimization (GiGPO), across two model families spanning 1.5B to 7B. At 7B, it reaches 96.1% ALFWorld success and a WebShop score of 93.1. Frozen-batch analyses show that ICSD reduces teacher-supported mass assigned to objective-opposed tokens from 60.1% to 37.8% and raises cosine compatibility with the RL gradient by 0.192. A companion repository is available at https://github.com/lanqz7766/Influence-Calibration-for-On-Policy-Self-Distillation-in-Agentic-RL.

Introduction

Language agents are increasingly trained with on-policy reinforcement learning (RL) for long-horizon interaction. The reward carries the task objective but arrives once per trajectory, far more coarsely than the sequence of decisions that produced it: a delayed outcome rates the whole episode even when a few actions decide success. Recent work closes this granularity gap with dense auxiliary supervision from a privileged reference (Wang et al. 2026; Lu et al. 2026). The density has a price. An auxiliary distillation loss records what the reference prefers at each token, but by itself does not reveal whether moving toward that preference helps the objective currently being optimized.

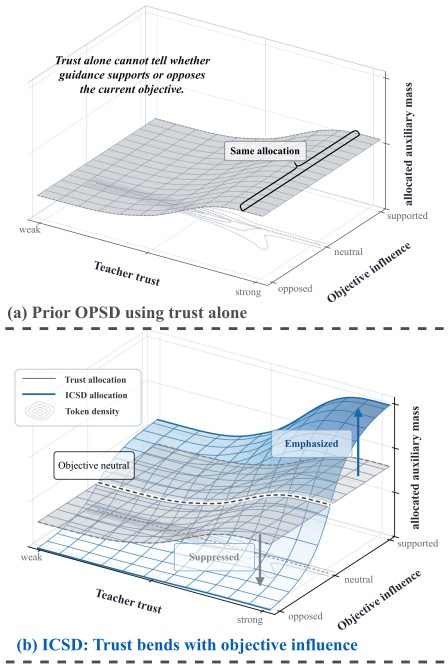

Figure 1: Many existing OPSD allocation rules determine token weights using teacher trust alone. (a) Tokens with the same trust therefore receive the same coefficient, even when their influence on the RL objective differs. (b) ICSD retains the trust signal and uses objective influence to redistribute the same auxiliary mass.

On-policy self-distillation (OPSD) supplies exactly this kind of dense signal: it evaluates a teacher or privileged branch on the student's own rollouts and supervises tokens at states the current policy visits (Zhao et al. 2026). Skill-conditioned variants extend this idea to multi-turn agents by exposing the teacher, but not the student, to trajectory-derived

*These authors contributed equally.
†Corresponding author.

skills or other privileged context (Wang et al. 2026; Lu et al. 2026). This design reduces the distribution mismatch of offline teacher trajectories and turns sparse outcome feedback into token-level guidance. It also creates a different problem: once a rollout contains many teacher-supervised items, how should the auxiliary update be distributed among them?

Existing selection rules primarily estimate whether a teacher signal is informative or trustworthy from teacher–student agreement (Lu et al. 2026), uncertainty and disagreement (Xu et al. 2026b), token position (Liu et al. 2026), or outcome evidence (Akhondzadeh et al. 2026). Useful as these quantities are, they do not determine the utility of an auxiliary update for the current RL objective. Two sampled tokens can receive equally strong teacher support while one sits in a turn the advantage estimator credits and the other in a turn it penalizes. We call this gap the *trust–utility mismatch*: trust asks whether a correction is credible, utility asks whether moving toward it serves the current policy-improvement direction, and neither answers the other. On frozen ALFWorld batches, for example, trust-only allocation places 60.1% of its teacher-supported mass on tokens the current objective opposes (Figure 3). We therefore condition trusted supervision on objective utility rather than replacing the teacher-trust signal.

Turning this comparison into an allocation rule raises three challenges:

- *Influence estimation.* Classical influence functions answer how upweighting one item changes an objective, but their parameter-space evaluation needs inverse-curvature information and per-item gradients, prohibitive during online training of billion-parameter agents (Koh and Liang 2017).
- *Distribution calibration.* Local sensitivities inherit the drifting scale and heavy tails of advantages, ratios, and teacher gaps; a fixed threshold means different things across turns and training stages.
- *Signal validity.* A local signal is directional only where the deployed loss can realize the analyzed direction; elsewhere, sign agreement is not evidence, and the allocator must stay conservative.

We address these challenges with Influence Calibration for Self-Distillation (ICSD), an objective-informed allocation rule for the existing OPSD auxiliary branch. ICSD obtains its influence signal from a teacher-directed perturbation of the sampled output. The score is the first-order Taylor term of the token's local RL-surrogate response and uses only quantities already available from the policy update. A batch-adaptive map converts these drifting, heavy-tailed scores into bounded relative multipliers that remain comparable across turns and training stages. Where the local interpretation is ambiguous, ICSD retains the inherited SDAR allocation (Lu et al. 2026). Exact action-turn mass matching then ensures that influence redistributes trusted supervision without increasing its total amount.

Our contributions are:

- We identify the trust–utility mismatch in OPSD and formulate token-level distillation as objective-informed allocation of a detached auxiliary update.
- We derive a teacher-directed objective-influence score and turn it into a continuous allocation rule through batch-adaptive calibration, a conservative SDAR fallback, and exact action-turn mass matching, without an additional model pass.
- Across three agent benchmarks, two policy optimization algorithms, and five model configurations, ICSD yields consistent aggregate gains over trust-only allocation, while frozen-batch analyses verify the predicted reallocation from objective-opposed to objective-supported teacher corrections.

Related Work

Privileged and Selective On-Policy Distillation. OPSD evaluates a stop-gradient privileged self-teacher on student rollouts (Zhao et al. 2026). Skill-SD supplies trajectory-derived skills to the privileged branch, SDAR gates its loss by the teacher–student gap, and SAGE-OPD and TurnOPD account for where supervision occurs in a long interaction (Wang et al. 2026; Lu et al. 2026; Zhou et al. 2026a,b). Other selectors estimate trust from entropy and divergence, token position or prefix discrepancy, Bayesian evidence, or contrasted privileged hints (Xu et al. 2026b; Liu et al. 2026; Xie et al. 2026; Shen et al. 2026; Pan et al. 2026); privileged supervision can also erase high-entropy forks and self-correction (Kaur et al. 2026). These methods determine whether and where teacher evidence is trustworthy. ICSD retains that signal but asks whether the resulting token update supports the local RL objective.

Coupling Distillation to the RL Objective. RL feedback enters prior distillation methods through several interfaces. RLSD, EGRSD, and PBSD modify RL advantages (Yang et al. 2026; Ke et al. 2026; Tian et al. 2026); SG-OPD uses a binary sequence-level gate, DOPD changes the supervision source and form, RG-OPD filters trajectories by reward–teacher agreement, and CRAFT adds an actor loss from sibling-rollout credit (Xu et al. 2026a; Yu et al. 2026; Akhondzadeh et al. 2026; Meng and Chen 2026). ICSD instead leaves the RL surrogate unchanged and uses its existing advantage and ratio only to allocate the detached OPSD loss per token. This output-coordinate view is related to gradient balancing (Yu et al. 2020; Chen et al. 2018; Du et al. 2018), while influence functions motivate the item-upweighting question and Fisher geometry its idealized parameter-space interpretation (Koh and Liang 2017; Schulman et al. 2015).

Method

ICSD is defined directly on the privileged-distillation objective. Starting from its token-level form, we derive objective influence and use it to construct calibrated coefficients that preserve the inherited auxiliary mass. Figure 2 summarizes this allocation pipeline from privileged token scoring to the ICSD-weighted actor update.

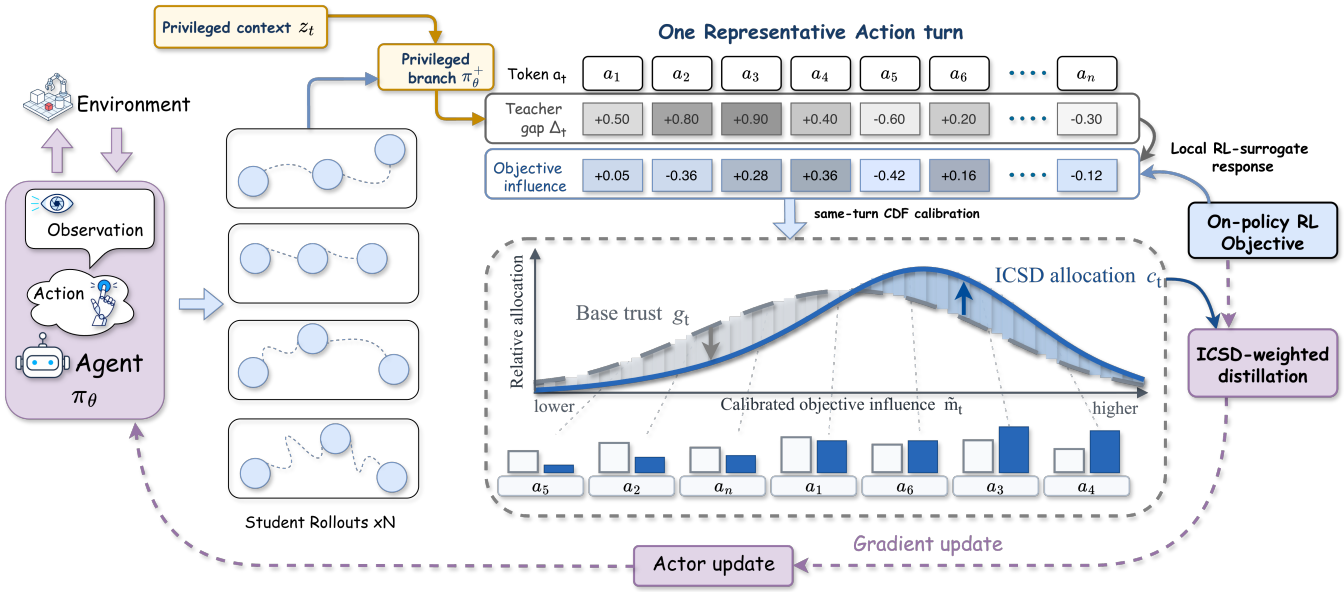

Figure 2: Overview of ICSD. A privileged branch evaluates the student's on-policy action tokens, while the RL objective provides token-level objective influence. ICSD calibrates this influence and uses it to redistribute the inherited trust allocation before the actor update.

On-Policy Privileged Distillation

Let $\mathcal{B}$ be the set of valid response-token positions in an on-policy minibatch. At position $t \in \mathcal{B}$, the student π_θ observes context x_t and samples token a_t. The privileged branch $\pi_\theta^+(\cdot \mid x_t, z_t)$ evaluates that token with additional context z_t, such as a trajectory-derived skill, that is withheld from the student. Both branches share the evolving actor parameters. During the current actor update, however, the privileged log probability is treated as a constant. The per-token auxiliary loss is

$$\hat{\ell}_t^{\mathrm{KD}}(\theta) = \mathrm{sg}\big[\log \pi_\theta^+(a_t \mid x_t, z_t)\big] - \log \pi_\theta(a_t \mid x_t), \quad (1)$$

where sg denotes stop-gradient. We also record the detached teacher–student gap

$$\Delta_t = \mathrm{sg}\big[\log \pi_\theta^+(a_t \mid x_t, z_t) - \log \pi_\theta(a_t \mid x_t)\big], \quad (2)$$

A positive Δ_t means that the privileged branch assigns greater probability to the sampled token than the student does.

Uniform OPSD averages these token losses equally. More generally, let $c_t \geq 0$ be a detached coefficient that controls how much auxiliary weight token t receives:

$$\mathcal{L}(\theta) = \mathcal{L}_{\mathrm{RL}}(\theta) + \frac{\lambda}{|\mathcal{B}|} \sum_{t \in \mathcal{B}} \mathrm{sg}(c_t)\, \hat{\ell}_t^{\mathrm{KD}}(\theta). \quad (3)$$

Here λ is the distillation weight and $|\mathcal{B}|$ is the number of valid response tokens in the minibatch. SDAR derives c_t from Δ_t, treating the sampled gap as evidence of whether the teacher signal should be trusted (Lu et al. 2026). We denote this inherited trust weight by $g_t = G_{\mathrm{trust}}(\Delta_t) \in [0, 1]$, where G_{trust} is monotone. In our experiments, G_{trust} is the same asymmetric Laplace CDF defined below, applied to Δ_t with split parameter 0.4. ICSD leaves this SDAR component unchanged. Because g_t depends only on Δ_t, two tokens with the same gap receive the same trust weight. Their relation to the current RL objective is not represented.

Teacher-Directed Objective Influence

To add the missing objective-side information, we measure how the current RL surrogate responds when the privileged branch perturbs the sampled token. Let π_{old} be the rollout policy held fixed during the actor update. The policy ratio is $\rho_t = \pi_\theta(a_t \mid x_t)/\pi_{\mathrm{old}}(a_t \mid x_t)$, and $\widehat{A}_t$ is the token advantage supplied by the policy optimization algorithm. Token t contributes $J_t = \widehat{A}_t \rho_t$ to the unclipped maximization surrogate.

Classical parameter-space influence requires applying inverse curvature to a per-item parameter gradient, typically through a linear solve or repeated Hessian–vector products (Koh and Liang 2017). Repeating that computation for every supervised token inside each actor update is impractical for billion-parameter agents. We instead perturb the scalar output coordinate specified by the teacher–student gap:

$$\log \rho_t(\epsilon) = \log \rho_t + \epsilon \Delta_t, \qquad \rho_t(\epsilon) = \rho_t \exp(\epsilon \Delta_t). \quad (4)$$

The resulting teacher-directed objective influence is

$$u_t := \left. \frac{\mathrm{d}\, \widehat{A}_t \rho_t(\epsilon)}{\mathrm{d}\epsilon} \right|_{\epsilon=0} = \widehat{A}_t \rho_t \Delta_t, \quad (5)$$

$$J_t(\epsilon) = J_t(0) + \epsilon u_t + \mathcal{O}(\epsilon^2). \quad (6)$$

Thus, u_t is the exact first-order Taylor coefficient for the teacher-directed output-coordinate intervention in Equation (4). For teacher-supported tokens ($\Delta_t > 0$), its sign

indicates whether increasing the sampled-token probability raises or lowers the local surrogate contribution. Its magnitude varies with the advantage and policy ratio, so it must be calibrated before being used for allocation.

Calibrated Influence Allocation

Turn-conditioned calibration. The absolute scale of u_t is not stable across minibatches because it inherits the scale of the advantage and policy ratio. Its product form can also yield heavy tails and unequal dispersion below and above the batch center. We therefore use the median to limit outlier leverage and fit separate scales on the two sides. Recall that $\mathcal{B}$ contains all valid response-token positions in the current minibatch. An action turn is the span emitted for one agent action in one trajectory. For each turn index, the calibration group $\mathcal{G} \subseteq \mathcal{B}$ pools valid tokens at that index across trajectories. Sparse groups use the statistics of all tokens in $\mathcal{B}$.

Calibration map. For each $\mathcal{G}$, we estimate a location and two one-sided scales:

$$\begin{aligned} \hat{\mu}_{\mathcal{G}} &= \mathrm{median}_{j\in\mathcal{G}}\, u_j, \\ \hat{b}^{-}_{\mathcal{G}} &= \mathrm{mean}_{j\in\mathcal{G}:u_j<\hat{\mu}_{\mathcal{G}}} \left|u_j - \hat{\mu}_{\mathcal{G}}\right|, \\ \hat{b}^{+}_{\mathcal{G}} &= \mathrm{mean}_{j\in\mathcal{G}:u_j\geq\hat{\mu}_{\mathcal{G}}} \left|u_j - \hat{\mu}_{\mathcal{G}}\right|. \end{aligned} \tag{7}$$

An empty side is assigned the scale ε, and every fitted scale is floored at the same small $\varepsilon > 0$. The calibrated map is

$$\Phi_{\mathcal{G}}(u) = \begin{cases} \eta \exp\Big((u-\hat{\mu}_{\mathcal{G}})/\hat{b}^{-}_{\mathcal{G}}\Big), & u < \hat{\mu}_{\mathcal{G}}, \\ 1-(1-\eta)\exp\Big(-(u-\hat{\mu}_{\mathcal{G}})/\hat{b}^{+}_{\mathcal{G}}\Big), & u \geq \hat{\mu}_{\mathcal{G}}, \end{cases} \tag{8}$$

Equation (8) is an asymmetric Laplace-family CDF with split parameter $\eta = 0.5$. We set $\tilde{m}_t = \Phi_{\mathcal{G}}(u_t) \in (0,1)$. It is a bounded relative score within $\mathcal{G}$, not an event probability.

Conservative composition. The calibrated score is not meaningful under every sign pattern. The auxiliary loss in Equation (1) raises the sampled log probability, while the intervention in Equation (4) points downward when $\Delta_t < 0$. If $\widehat{A}_t < 0$ as well, their product makes u_t positive even though the two directions disagree. We retain the inherited SDAR coefficient in this case. Let $\mathcal{D}$ denote this fallback set and let $\mathcal{C}$ denote the remaining tokens:

$$\mathcal{D} = \{t \in \mathcal{B} : \widehat{A}_t < 0,\ \Delta_t < 0\}, \qquad \mathcal{C} = \mathcal{B} \setminus \mathcal{D}. \tag{9}$$

For mass matching, let $q \subseteq \mathcal{B}$ denote one action turn and let $\mathcal{Q}$ collect all turns in the minibatch. These turns form a disjoint partition, $\mathcal{B} = \biguplus_{q\in\mathcal{Q}} q$. For each $q \in \mathcal{Q}$, write $\mathcal{C}_q = \mathcal{C} \cap q$ and $\mathcal{D}_q = \mathcal{D} \cap q$. Thus, $\mathcal{G}$ pools a turn index across trajectories for calibration, whereas q refers to one specific turn for mass matching. When $\sum_{j\in\mathcal{C}_q} g_j \tilde{m}_j > 0$, we set

$$\alpha_q = \frac{\sum_{j\in\mathcal{C}_q} g_j}{\sum_{j\in\mathcal{C}_q} g_j \tilde{m}_j}, \tag{10}$$

$$c_t = \begin{cases} g_t, & t \in \mathcal{D}_q, \\ \alpha_q g_t \tilde{m}_t, & t \in \mathcal{C}_q. \end{cases} \tag{11}$$

If the denominator is numerically zero, we use $c_t = g_t$ throughout that turn. The resulting allocation has the following properties.

Proposition 1 (Invariant mass allocation) *Assume nondegenerate fitted scales and $\eta \in (0,1)$. For fixed $\mathcal{G}$, $\Phi_{\mathcal{G}}$ is monotone in u. Consider a positive affine transformation $u'_j = au_j + b$, $a > 0$, applied to every score used to fit a calibration group, while holding the trust weights and fallback partition fixed. Then $\tilde{m}'_j = \tilde{m}_j$. When the denominator in Equation (10) is nonzero, $\alpha'_q = \alpha_q$, and in either case $c'_t = c_t$. The coefficients also satisfy*

$$\sum_{t\in q} c_t = \sum_{t\in q} g_t \qquad \text{for every } q \in \mathcal{Q}. \tag{12}$$

For $t, t' \in \mathcal{C}_q$ with $g_t = g_{t'}$, $u_t \leq u_{t'}$ implies $c_t \leq c_{t'}$.

Proof sketch. The two branches of Equation (8) have positive derivatives and meet at η, so $\Phi_{\mathcal{G}}$ is monotone. A positive affine transform preserves each score's side of the median. The fitted location and one-sided scales transform as $a\hat{\mu}_{\mathcal{G}} + b$ and $a\hat{b}^{\pm}_{\mathcal{G}}$, leaving $\tilde{m}_t$ unchanged. With the trust weights and fallback partition fixed, the sums defining α_q are also unchanged. The same is therefore true of α_q and c_t. Expanding the two parts of each turn gives Equation (12), and equal-trust tokens inherit the ordering of $\tilde{m}_t$. Supplementary Section A gives the full derivation.

Training Objective

The final actor update uses Equation (3) with the coefficient from Equation (11). Both c_t and the privileged log probability inside $\hat{\ell}^{\mathrm{KD}}_t$ are detached, so auxiliary gradients pass only through the student log probability. Equation (12) preserves SDAR's coefficient mass within each action turn, and computing c_t requires no additional model evaluation. Supplementary Section C reports the measured overhead.

Experiments and Results

Setup

We evaluate language agents on ALFWorld (Shridhar et al. 2020), WebShop (Yao et al. 2022), and the seven Search-QA subsets used by Search-R1 (Jin et al. 2025). We use the task splits and metrics of SDAR (Lu et al. 2026). Experiments cover Qwen2.5-Instruct models at 1.5B, 3B, and 7B (Yang et al. 2024), as well as Qwen3-1.7B and Qwen3-4B (Yang et al. 2025). We evaluate ICSD with both GRPO (Shao et al. 2024) and GiGPO (Feng et al. 2026).

Within each matched comparison, the model, data, RL algorithm, privileged self-teacher, and inherited SDAR trust allocation are fixed. Published baseline rows retain the values reported by SDAR; the remaining entries come from our training and evaluation logs. Supplementary Table A1 summarizes the shared SDAR-aligned training protocol.

Main Results

Table 1 provides the matched comparison. Replacing SDAR's trust-only allocation with ICSD improves every summary metric in the GiGPO blocks, and the 1.5B GRPO

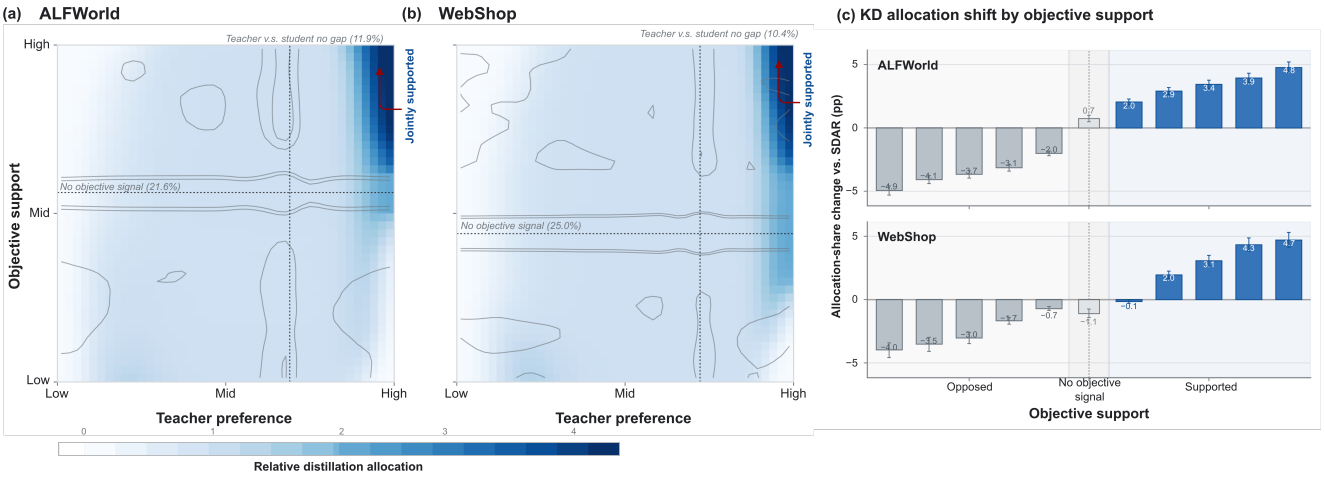

Figure 3: Continuous allocation induced by ICSD on frozen GiGPO batches. Panels (a–b) show normalized coefficients over teacher-preference and objective-support ranks; darker blue denotes more mass, contours show token density, and dotted lines mark zero signals. Panel (c) reports the shift relative to SDAR among teacher-supported tokens. Whiskers are 95% action-turn bootstrap intervals over four batches.

control shows the same ordering. The Qwen2.5-7B result illustrates the scale of the difference: ALFWorld success rises from 94.5 to 96.1, while WebShop score/accuracy moves from 88.4/78.9 to 93.1/84.4.

Table 2 broadens the comparison to Vanilla and Skill-Prompt, stand-alone OPSD (Zhao et al. 2026), and the Skill-GRPO variants used by SDAR (Lu et al. 2026). GiGPO+ICSD gives the strongest ALFWorld and Search-QA averages and the highest WebShop accuracy. Skill-Prompt* and Skill-GRPO* retain retrieved skills during evaluation, whereas ICSD requires no privileged context at test time.

Ablations and Analysis

Across policy optimizers. To test whether the benefit depends on the policy optimizer, we hold the model and task settings fixed at Qwen2.5-1.5B and compare the full training trajectories under GRPO and GiGPO (Supplementary Figure A3). ICSD outperforms the matched SDAR allocator on ALFWorld and WebShop with either optimizer. The gap is largest on WebShop accuracy: +11.7 points with GRPO and +7.0 with GiGPO. The result is therefore not specific to GiGPO's step-aware advantage construction.

Qwen3 model family. Table 3 repeats the matched comparison on Qwen3-4B-Instruct (Yang et al. 2025). With GiGPO and the concise non-thinking action interface fixed, ICSD improves average ALFWorld success from 89.8 to 95.3 over SDAR. This 5.5-point gain complements the Qwen2.5 results without changing the allocation rule.

Allocation ablation. Table 4 separates three parts of the allocation rule: signed objective relevance, teacher trust, and continuous influence magnitude. The Fisher-magnitude control passes an unsigned policy-gradient magnitude through the same CDF calibration; the teacher gap appears only in the inherited SDAR trust weight. Its 91.4% average matches GiGPO, showing that sensitivity magnitude alone is not

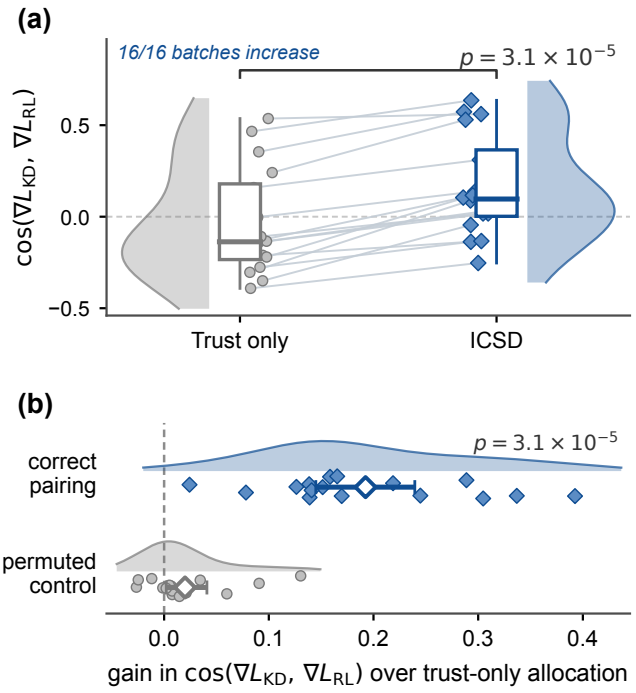

Figure 4: ICSD improves local update compatibility on sixteen frozen batches. (a) SDAR and ICSD cosine compatibility with the RL gradient. (b) Compatibility after tokenwise permutation of the ICSD multipliers. Diamonds and horizontal bars show means and 95% batch-bootstrap intervals.

enough. Influence-only allocation instead retains the signed score but removes token-level trust, raising the average to 93.0%. Combining both signals gives the full 93.8% result.

The sign-only control replaces the continuous calibrated score with a hard positive-or-negative decision while retain-

	ALFWorld							Search-QA								WebShop	
Method	Pick	Look	Clean	Heat	Cool	Pick2	Avg	NQ	Triv	Pop	Hotp	2Wk	MuS	Bam	Avg	Score	Acc.
Qwen2.5-1.5B-Instruct																	
Skill-SD	88.9	57.1	60.0	80.0	65.0	43.5	68.0	16.7	35.4	14.8	16.7	25.2	3.9	11.2	17.7	80.6	68.8
RLSD	78.1	41.7	68.2	64.3	77.3	57.7	67.2	13.9	29.4	12.4	14.2	21.1	3.0	9.5	14.8	80.5	63.3
GRPO	84.4	41.7	77.3	50.0	68.2	50.0	65.6	16.6	34.1	15.9	16.1	24.8	3.4	58.1	24.1	72.0	60.9
+ OPSD	90.6	33.3	77.3	57.1	77.3	61.5	71.1	16.9	34.3	16.4	16.3	24.8	3.4	56.9	24.2	84.9	73.4
+ SDAR	87.1	64.7	75.0	66.7	50.0	80.0	72.7	16.1	32.7	15.8	15.9	24.4	3.3	4.4	16.1	81.9	64.1
+ ICSD	87.5	71.4	81.5	62.5	75.8	52.4	73.4	18.3	39.2	17.1	18.7	28.2	4.4	68.0	**27.7**	85.6	75.8
GiGPO	97.1	75.0	95.8	88.9	84.2	91.7	91.4	16.8	35.3	14.6	17.1	24.5	3.6	12.8	17.8	83.9	73.4
+ SDAR	97.2	88.9	100.0	94.1	95.0	66.7	92.2	16.5	34.4	14.4	16.7	24.7	3.1	11.2	17.3	88.8	71.1
+ ICSD	93.1	91.7	92.0	100.0	95.5	92.3	**93.8**	17.5	36.8	14.5	17.7	24.4	4.1	12.7	18.2	**92.2**	**78.1**
Qwen2.5-3B-Instruct																	
Skill-SD	88.2	50.0	96.2	52.4	65.0	57.9	73.4	44.4	60.4	44.0	39.5	40.4	15.4	64.9	44.1	75.9	64.0
RLSD	87.9	75.0	90.9	75.0	73.1	68.4	79.7	41.5	58.6	42.3	40.4	40.2	16.8	66.9	43.8	84.4	66.4
GRPO	91.2	62.5	96.2	61.9	65.0	47.4	75.0	39.3	60.6	41.1	37.4	34.6	15.4	26.4	36.4	79.8	63.3
+ OPSD	100.0	82.4	85.7	75.0	70.0	60.0	81.2	44.9	61.2	45.2	40.4	38.5	16.0	66.1	**44.6**	77.8	66.4
+ SDAR	97.1	62.5	100.0	61.9	75.0	84.2	84.4	44.8	58.1	44.3	38.6	36.2	15.7	66.1	43.4	85.0	68.0
GiGPO	100.0	75.0	95.5	78.6	86.4	100.0	92.2	42.0	59.5	42.4	36.9	37.0	12.6	64.1	42.1	86.3	75.0
+ SDAR	100.0	83.3	96.6	100.0	79.2	91.3	93.0	44.7	60.0	44.6	37.6	32.9	13.9	64.5	42.6	87.2	74.2
+ ICSD	100.0	92.9	100.0	86.7	95.0	91.3	**95.3**	43.0	60.3	43.9	38.2	40.0	13.4	64.5	43.3	**88.7**	**75.8**
Qwen2.5-7B-Instruct																	
Skill-SD	93.9	93.8	90.9	100.0	69.2	68.4	85.1	47.1	64.5	47.8	44.2	42.1	20.2	69.0	47.8	86.1	76.5
RLSD	100.0	87.5	92.3	58.8	80.0	65.2	82.0	46.8	63.0	44.4	45.5	48.9	21.5	73.0	49.0	87.4	77.3
GRPO	91.2	87.5	96.2	81.0	65.0	57.9	81.2	45.1	63.7	44.0	43.6	43.2	16.8	37.6	42.0	80.9	72.6
+ OPSD	91.4	61.5	100.0	87.5	76.5	52.2	80.4	47.3	64.5	46.9	43.8	39.3	18.0	69.4	47.0	86.8	76.5
+ SDAR	94.7	75.0	100.0	86.7	68.2	78.9	85.9	46.3	63.5	48.2	43.8	48.4	19.6	73.0	49.0	89.4	82.8
GiGPO	100.0	58.3	90.9	100.0	81.8	100.0	91.4	46.4	64.7	46.1	41.6	43.6	18.9	68.9	47.2	84.0	71.1
+ SDAR	100.0	87.0	100.0	87.5	90.9	95.2	94.5	45.1	63.8	46.3	41.8	41.8	19.4	72.2	47.2	88.4	78.9
+ ICSD	100.0	91.7	90.9	100.0	95.5	96.2	**96.1**	47.8	64.6	48.5	44.2	45.3	20.1	73.3	**49.1**	**93.1**	**84.4**
Qwen3-1.7B-Instruct																	
Skill-SD	52.9	37.5	69.2	42.9	60.0	36.8	52.3	39.1	57.5	45.4	34.8	34.1	10.7	64.1	40.8	**81.8**	53.9
RLSD	50.0	37.5	61.5	19.0	50.0	21.1	42.2	38.6	57.3	43.0	34.5	34.1	11.5	65.3	40.6	74.0	50.8
GRPO	71.1	41.7	36.4	40.0	31.8	31.6	46.1	40.0	58.9	43.5	35.4	30.3	12.0	65.7	40.8	67.3	38.3
+ OPSD	38.2	50.0	30.8	28.6	30.0	21.1	32.0	40.7	58.9	45.0	37.0	34.6	13.3	65.7	42.2	70.7	38.3
+ SDAR	73.5	25.0	76.9	33.3	40.0	36.8	53.9	39.7	58.9	45.3	35.9	35.5	12.6	65.3	41.9	76.8	58.6
GiGPO	85.3	50.0	96.2	66.7	75.0	68.4	78.1	43.1	58.4	44.8	33.8	28.9	10.9	62.9	40.4	79.9	60.2
+ SDAR	72.7	62.5	90.9	58.3	84.6	68.4	75.0	42.1	59.1	46.7	34.8	29.7	11.0	63.7	41.0	78.4	65.6
+ ICSD	94.1	62.5	100.0	71.4	85.0	57.9	**82.8**	43.6	59.7	47.0	36.3	31.3	13.2	66.1	**42.5**	81.5	**68.0**

Table 1: Main results grouped by policy optimizer. Skill-SD and RLSD are hybrid distillation baselines; family-head rows are the RL baselines, "+" rows add auxiliary allocators, and blue rows denote ICSD. All values are percentages. Within each model block, bold and underline mark the best and second-best summary results; task-level columns are unmarked.

Method	ALFWorld Avg	Search-QA Avg	WebShop Score	WebShop Acc.
Qwen2.5-1.5B-Instruct				
Vanilla	5.5	12.1	17.8	5.5
Skill-Prompt*	6.2	7.7	20.8	1.6
OPSD	14.1	0.0	22.3	10.2
GiGPO+ICSD	**93.8**	**18.2**	**92.2**	**78.1**
Qwen2.5-3B-Instruct				
Vanilla	21.9	31.7	6.7	0.8
Skill-Prompt*	28.9	23.9	0.2	0.8
OPSD	28.1	0.0	11.3	3.1
Skill-GRPO	60.2	34.1	77.3	60.9
Skill-GRPO*	80.5	36.1	76.3	66.4
GiGPO+ICSD	**95.3**	**43.3**	**88.7**	**75.8**
Qwen2.5-7B-Instruct				
Vanilla	12.5	33.9	5.9	1.6
Skill-Prompt*	23.4	36.4	1.7	0.8
OPSD	32.8	6.2	4.5	2.3
Skill-GRPO	69.5	40.3	80.4	71.9
Skill-GRPO*	88.3	47.5	87.0	81.2
GiGPO+ICSD	**96.1**	**49.1**	**93.1**	**84.4**
Qwen3-1.7B-Instruct				
Vanilla	12.5	24.8	46.5	4.7
Skill-Prompt*	9.4	24.3	23.0	2.3
OPSD	14.1	5.8	47.4	9.3
Skill-GRPO	21.1	40.4	73.4	46.1
Skill-GRPO*	28.1	40.7	80.4	50.0
GiGPO+ICSD	**82.8**	**42.5**	**81.5**	**68.0**

Table 2: Summary results for extended baselines. Starred methods retain retrieved skills during validation. Supplementary Table A2 reports per-task and per-subset results.

Method	Pick	Look	Clean	Heat	Cool	Pick2	Avg
GiGPO+SDAR	96.9	66.7	**100.0**	78.6	81.8	**96.2**	89.8
GiGPO+ICSD	**97.6**	**91.7**	**100.0**	**88.9**	**100.0**	83.3	**95.3**

Table 3: Qwen3-4B ALFWorld results with GiGPO and the concise non-thinking interface.

Method	Pick	Look	Clean	Heat	Cool	Pick2	Avg
GiGPO	97.1	75.0	95.8	88.9	84.2	91.7	91.4
+ SDAR	97.2	88.9	**100.0**	94.1	95.0	66.7	92.2
+ Fisher magnitude	**100.0**	78.6	95.0	**100.0**	90.0	78.3	91.4
+ Sign-only	97.1	**100.0**	90.9	**100.0**	**95.5**	77.3	92.2
+ Influence only	93.9	88.9	92.3	**100.0**	**95.5**	83.3	93.0
+ ICSD	93.1	**91.7**	92.0	**100.0**	**95.5**	**92.3**	**93.8**

Table 4: Allocation ablation on ALFWorld with Qwen2.5-1.5B and GiGPO.

ing the fallback and exact turn-wise mass matching. It reaches 92.2%, tying SDAR but remaining 1.6 points below ICSD. Thus, the gain is not reproduced by unsigned sensitivity or hard sign filtering alone; the strongest result comes from combining teacher trust with graded, signed objective relevance. Supplementary Figures A1–A2 visualize fallback

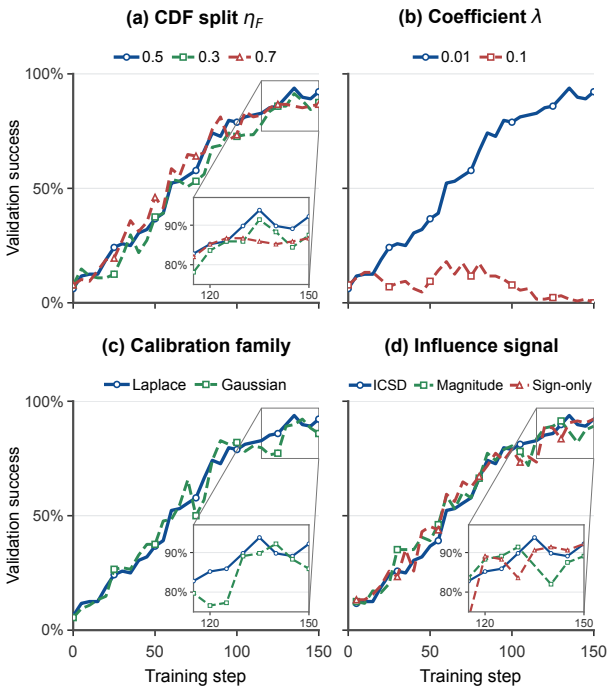

Figure 5: Training dynamics for sensitivity and signal ablations on ALFWorld with Qwen2.5-1.5B and GiGPO. Insets enlarge steps 115–150.

behavior and selected validation trajectories.

Sensitivity. Figure 5 motivates the settings used in our main experiments. The central CDF split and $\lambda = 0.01$ give the strongest late-stage performance in the tested ranges, and asymmetric Laplace calibration finishes above the Gaussian alternative. Continuous signed influence also reaches a higher peak and late-window mean than magnitude-only or sign-only allocation.

Mechanism analysis. For nonnegative token weights w_t, trusted-conflict mass (TCM) is the fraction of active teacher-supported mass ($\Delta_t > 0$, $|u_t| > \varepsilon$) assigned to objective-opposed tokens ($u_t < 0$). The exact estimator and four-region breakdown are provided in Supplementary Section D and Figure A1. Across four matched ALFWorld batches, TCM falls from 60.1% under SDAR to 37.8% under ICSD, a 22.3-point reduction (95% CI $[22.1, 22.6]$). Figure 3 shows the same redistribution continuously: strongly trusted and supported regions receive $2.90\times$ and $2.64\times$ mass on ALFWorld and WebShop, while trusted but opposed ALFWorld tokens receive $0.79\times$.

Across sixteen frozen batches, ICSD raises cosine compatibility with the RL gradient by 0.192 (95% CI $[0.147, 0.240]$), whereas permuting the same multipliers leaves only a 0.020 gain (Figure 4). The compatibility gain is positive in all sixteen batches; Supplementary Section D reports the paired bootstrap analysis. The improvement therefore comes from token assignment rather than total mass.

Conclusion

ICSD addresses the trust–utility mismatch in on-policy self-distillation by reallocating trusted token supervision according to teacher-directed objective influence. Batch-adaptive calibration, conservative fallback, and exact action-turn mass matching keep the RL objective and auxiliary budget unchanged. Across three benchmarks, two optimizers, and five model configurations, ICSD consistently improves over trust-only allocation; frozen-batch analyses show that it shifts mass toward objective-supported corrections and improves local update compatibility.

References

Akhondzadeh, M. S.; Lingam, V.; Tejaswi, A.; Ekbote, C.; Sanghavi, S.; and Bojchevski, A. 2026. Reward-Gated On-Policy Distillation. *arXiv preprint arXiv:2607.04037*.

Chen, Z.; Badrinarayanan, V.; Lee, C.-Y.; and Rabinovich, A. 2018. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, 794–803. PMLR.

Du, Y.; Czarnecki, W. M.; Jayakumar, S. M.; Farajtabar, M.; Pascanu, R.; and Lakshminarayanan, B. 2018. Adapting auxiliary losses using gradient similarity. *arXiv preprint arXiv:1812.02224*.

Feng, L.; Xue, Z.; Liu, T.; and An, B. 2026. Group-in-group policy optimization for llm agent training. *Advances in Neural Information Processing Systems*, 38: 46375–46408.

Jin, B.; Zeng, H.; Yue, Z.; Yoon, J.; Arik, S.; Wang, D.; Zamani, H.; and Han, J. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.

Kaur, S.; Ri, N.; He, Y.; Fowl, L. H.; and Arora, S. 2026. Rethinking On-Policy Self-Distillation for Thinking Models. In *Workshop on Failure Modes of Agentic AI at ICML 2026*.

Ke, J.; Wen, Z.; Li, W.; He, C.; and Zhang, L. 2026. Respecting self-uncertainty in on-policy self-distillation for efficient llm reasoning. *arXiv preprint arXiv:2605.13255*.

Koh, P. W.; and Liang, P. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*, 1885–1894. PMLR.

Liu, X.; Wang, X.; Ma, Y.; Zhang, Y.; and Xiao, C. 2026. When Are Teacher Tokens Reliable? Position-Weighted On-Policy Self-Distillation for Reasoning. *arXiv preprint arXiv:2605.21606*.

Lu, Z.; Yao, Z.; Han, Z.; Wang, Z.-H.; Wu, J.; Gu, Q.; Cai, X.; Lu, W.; Xiao, J.; Zhuang, Y.; et al. 2026. Self-distilled agentic reinforcement learning. *arXiv preprint arXiv:2605.15155*.

Meng, Z.; and Chen, K. 2026. CRAFT: Counterfactual Credit Assignment from Free Sibling Rollouts for Self-Distilled Agentic Reinforcement Learning. *arXiv preprint arXiv:2606.29476*.

Pan, L.; Tao, S.; Zhai, Y.; Zhang, L.; Liu, Z.; Ding, B.; Liu, A.; and Wen, L. 2026. RLCSD: Reinforcement Learning with Contrastive On-Policy Self-Distillation. *arXiv preprint arXiv:2606.11709*.

Schulman, J.; Levine, S.; Abbeel, P.; Jordan, M.; and Moritz, P. 2015. Trust region policy optimization. In *International conference on machine learning*, 1889–1897. PMLR.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Shen, G.; Huang, L.; Cheng, X.; Zhao, C.; Li, J.; Zhao, D.; and Yu, X. 2026. From Generic Correlation to Input-Specific Credit in On-Policy Self Distillation. *arXiv preprint arXiv:2605.11613*.

Shridhar, M.; Yuan, X.; Côté, M.-A.; Bisk, Y.; Trischler, A.; and Hausknecht, M. 2020. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*.

Tian, Y.; Wang, R.; Wen, X.; Li, J.; Sun, S.; Song, L.; Bian, J.; and Zhao, B. 2026. PBSD: Privileged Bayesian Self-Distillation for Long-Horizon Credit Assignment. *arXiv preprint arXiv:2606.09348*.

Wang, H.; Wang, G.; Xiao, H.; Zhou, Y.; Pan, Y.; Wang, J.; Xu, K.; Wen, Y.; Ruan, X.; Chen, X.; and Qi, H. 2026. Skill-sd: Skill-conditioned self-distillation for multi-turn llm agents. *arXiv preprint arXiv:2604.10674*.

Xie, Y.; Zhu, S.; Wen, T.; Chen, B.; and Wang, Y. 2026. On the Position Bias of On-Policy Distillation. *arXiv preprint arXiv:2606.22600*.

Xu, H.; Wang, H.; Gao, Y.; Li, J.; Zhang, X.; and Yuan, X. 2026a. SG-OPD: Sign-Gated On-Policy Distillation via Sign-Consistency Gating and Phased Teacher Sampling. *arXiv preprint arXiv:2606.09304*.

Xu, Y.; Sang, H.; Zhou, Z.; He, R.; Wang, Z.; and Geramifard, A. 2026b. Tip: Token importance in on-policy distillation. *arXiv preprint arXiv:2604.14084*.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.

Yang, C.; Qin, C.; Si, Q.; Chen, M.; Gu, N.; Yao, D.; Lin, Z.; Wang, W.; Wang, J.; and Duan, N. 2026. Self-distilled rlvr. *arXiv preprint arXiv:2604.03128*.

Yao, S.; Chen, H.; Yang, J.; and Narasimhan, K. 2022. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35: 20744–20757.

Yu, T.; Kumar, S.; Gupta, A.; Levine, S.; Hausman, K.; and Finn, C. 2020. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836.

Yu, X.; Li, G.; Si, Q.; Zhang, G.; Xu, Y.; Wang, C.; Dong, S.; Tuo, K.; Zeng, X.; Feng, K.; et al. 2026. DOPD: Dual On-policy Distillation. *arXiv preprint arXiv:2606.30626*.

Zhao, S.; Xie, Z.; Liu, M.; Huang, J.; Pang, G.; Chen, F.; and Grover, A. 2026. Self-Distilled Reasoner: On-Policy

Self-Distillation for Large Language Models. *arXiv preprint arXiv:2601.18734*.

Zhou, Y.; Zhang, L.; Wu, Y.; Wang, M.; Peng, B.; Liu, J.; Fan, X.; and Zhao, Z. 2026a. SAGE-OPD: Selective Agent-Guided Intervention for Multi-Turn On-Policy Distillation. *arXiv preprint arXiv:2606.19659*.

Zhou, Y.; Zheng, K.; Li, H.; Peng, D.; Xu, C.; and Chen, J. 2026b. TurnOPD: Making On-Policy Distillation Turn-Aware for Efficient Long-Horizon Agent Training. *arXiv preprint arXiv:2607.05804*.

A Derivations and Proofs

This section gives the explicit Taylor remainder for objective influence and proves the affine-invariance statement from the main paper.

Why not parameter-space influence. For a training item t, classical influence has the form $-\nabla_\theta J^\top H^{-1} \nabla_\theta \ell_t$, where H is the parameter Hessian (Koh and Liang 2017). Computing it for every supervised token requires per-token parameter gradients and an inverse-curvature solve, usually approximated with repeated Hessian–vector products. This cost would be incurred inside every actor update. The objective influence score used by ICSD instead differentiates a scalar output coordinate using quantities already produced by the RL update.

A.1 Taylor Remainder for Objective Influence

Fix a token t and abbreviate $A = \widehat{A}_t$, $\rho = \rho_t > 0$, and $\Delta = \Delta_t$; all three are constants of the current iterate, and Δ is detached by construction. The perturbed surrogate contribution is $J(\epsilon) = A\rho\, e^{\epsilon\Delta}$, which is smooth in ϵ. Differentiating at zero gives

$$J'(0) = A\rho\Delta = u_t, \tag{A1}$$

which is exactly the objective influence score u_t defined in the main paper. Applying Taylor's theorem with the Lagrange remainder to e^x at $x = \epsilon\Delta$, there exists ξ between 0 and $\epsilon\Delta$ with $e^{\epsilon\Delta} = 1 + \epsilon\Delta + \frac{1}{2}(\epsilon\Delta)^2 e^{\xi}$. Substituting,

$$J(\epsilon) - J(0) - \epsilon u_t = \frac{A\rho\Delta^2}{2}\,\epsilon^2 e^{\xi}, \tag{A2}$$

and hence, bounding e^{ξ} by $e^{|\epsilon\Delta|}$,

$$\big|J(\epsilon) - J(0) - \epsilon u_t\big| \le \frac{|A|\rho\Delta^2}{2}\,\epsilon^2 e^{|\epsilon\Delta|}. \tag{A3}$$

The $\mathcal{O}(\epsilon^2)$ term in the main paper's first-order expansion therefore carries the explicit constant $\frac{1}{2}|A|\rho\Delta^2 e^{|\epsilon\Delta|}$, and no approximation enters the first-order coefficient itself.

A.2 Monotonicity and Affine Invariance

For nondegenerate $\hat{b}_-$ and $\hat{b}_+$, differentiating the two branches of the asymmetric Laplace CDF defined in the main paper gives

$$\frac{\mathrm{d}\Phi_{\mathcal{G}}}{\mathrm{d}u} = \begin{cases} \Phi_{\mathcal{G}}(u)/\hat{b}_-, & u < \hat{\mu}, \\ \big(1 - \Phi_{\mathcal{G}}(u)\big)/\hat{b}_+, & u \ge \hat{\mu}. \end{cases} \tag{A4}$$

Both derivatives are positive, and the branches meet at $\Phi_{\mathcal{G}}(\hat{\mu}) = \eta$. The map is therefore continuous and strictly increasing.

Write $\mathcal{J}_- = \{j : u_j < \hat{\mu}\}$ and $\mathcal{J}_+ = \{j : u_j \ge \hat{\mu}\}$. Let $u'_j = a u_j + b$ with $a > 0$. The median is affine equivariant, so $\hat{\mu}' = a\hat{\mu} + b$. The side assignment is preserved because $u'_j < \hat{\mu}'$ if and only if $u_j < \hat{\mu}$. Hence, $\mathcal{J}'_\pm = \mathcal{J}_\pm$. Each one-sided scale transforms as

$$\hat{b}'_\pm = \mathrm{mean}_{j \in \mathcal{J}_\pm} \big|u'_j - \hat{\mu}'\big| = a\hat{b}_\pm, \tag{A5}$$

because $|u'_j - \hat{\mu}'| = a|u_j - \hat{\mu}|$. The standardized residual is therefore invariant:

$$\frac{u' - \hat{\mu}'}{\hat{b}'_{\sigma(u')}} = \frac{a(u - \hat{\mu})}{a\,\hat{b}_{\sigma(u)}} = \frac{u - \hat{\mu}}{\hat{b}_{\sigma(u)}}, \tag{A6}$$

where $\sigma(u)$ selects the side of the median. Substitution into the asymmetric Laplace CDF gives $\Phi'_{\mathcal{G}}(u') = \Phi_{\mathcal{G}}(u)$ and hence $\tilde{m}'_t = \tilde{m}_t$. With g_t, $\mathcal{C}_q$, and $\mathcal{D}_q$ fixed, both sums in the turnwise scale are unchanged. It follows that $\alpha'_q = \alpha_q$ whenever the denominator is nonzero and $c'_t = c_t$. If the denominator is zero, both constructions use $c_t = g_t$. The scale floor is only a numerical safeguard. A binding floor is the degenerate case excluded by the affine-invariance statement.

A.3 Turn-Level Conservation and Ordering

Fix an action turn q with $\sum_{j \in \mathcal{C}_q} g_j \tilde{m}_j > 0$. By the turnwise scale and conservative multiplier defined in the main paper,

$$\begin{aligned} \sum_{t \in q} c_t &= \sum_{t \in \mathcal{D}_q} g_t + \alpha_q \sum_{t \in \mathcal{C}_q} g_t \tilde{m}_t \\ &= \sum_{t \in \mathcal{D}_q} g_t + \sum_{t \in \mathcal{C}_q} g_t = \sum_{t \in q} g_t, \end{aligned} \tag{A7}$$

which is the exact turn-mass identity stated in the main paper. In the degenerate case, the rule sets $c_t = g_t$ throughout the turn, so the identity holds directly.

Summing over all turns gives the global coefficient-mass identity

$$\sum_{t \in \mathcal{B}} c_t = \sum_{t \in \mathcal{B}} g_t \qquad \text{for every minibatch}, \tag{A8}$$

so ICSD does not increase the total coefficient budget relative to the inherited trust allocation.

For two tokens $t, t' \in \mathcal{C}_q$ in the same turn,

$$\frac{c_t}{c_{t'}} = \frac{g_t \tilde{m}_t}{g_{t'} \tilde{m}_{t'}}, \tag{A9}$$

because the common factor α_q cancels. At equal trust, the within-turn allocation therefore orders tokens by $\tilde{m}_t = \Phi_{\mathcal{G}}(u_t)$, which is monotone in u_t. In particular, opposing signs of $\widehat{A}_t$ and Δ_t give $u_t < 0$ and a lower calibrated score than positive-influence items. The shared positive factor α_q preserves this ordering, though it does not require every negative-influence coefficient to be smaller than its base trust coefficient. This is the ordering used in the main paper's mechanism-analysis figure.

B Idealized Natural-Gradient Interpretation

This section relates the exact output-coordinate score u_t to an idealized parameter-space update. The argument motivates its sign; it is not part of the deployed algorithm.

Setting. Let $s_t = \nabla_\theta \log \pi_\theta(a_t \mid x_t)$ be the sampled-token score and let F denote the Fisher information of the policy on visited states, the local metric underlying trust-region policy optimization (Schulman et al. 2015). Consider the idealized teacher-directed step

$$\delta\theta = \eta F^{-1} \Delta_t s_t, \qquad \eta > 0. \tag{A10}$$

Assume that the unclipped surrogate is locally active at t, F is positive definite on the span of s_t, and cross-token score correlations under F^{-1} are neglected. A first-order expansion of the token's own contribution $J_t = \widehat{A}_t \rho_t$ gives

$$J_t(\theta+\delta\theta) - J_t(\theta) = \eta \kappa_t u_t + o(\eta), \qquad \kappa_t = s_t^\top F^{-1} s_t \geq 0. \tag{A11}$$

Thus, whenever $\kappa_t > 0$, the same-token local contribution changes with the sign of u_t. The leverage κ_t is token dependent, so this argument establishes local sign compatibility rather than preservation of the full score ranking. It neither models cross-token interactions nor guarantees finite-horizon return improvement.

Relation to the fallback. The sign interpretation inherits the support restriction of the teacher-directed intervention defined in the main paper. When $\Delta_t < 0$, the analyzed intervention and the sampled-token auxiliary update point in different directions. If $\widehat{A}_t < 0$ as well, their product makes u_t positive and can falsely resemble supportive evidence. The conservative multiplier avoids this ambiguity by retaining the base trust coefficient in that double-negative region. Opposing-sign cases still yield $u_t < 0$ and remain useful as relative conflict scores.

C Implementation Details

C.1 Training Protocol

Protocol details. We use the SDAR data splits, SkillBank, keyword-matching retrieval, and privileged self-teacher. GRPO and GiGPO retain their original advantage construction; the GiGPO runs use $\gamma = 0.95$ and unit step-advantage weight. The ICSD-specific trust and influence split parameters are 0.4 and 0.5, with a scale floor of 10^{-4} and a minimum calibration group of eight tokens. Sparse groups use minibatch statistics before reverting to the fixed monotone map. Model-dependent micro-batches and tensor parallelism are chosen to fit the corresponding model sizes.

C.2 Runtime and Approximation Audit

Compute overhead. ICSD adds no forward or backward pass; its cost is limited to detached elementwise arithmetic and CDF fitting inside the actor update. On matched Qwen2.5-1.5B WebShop runs using four H100 GPUs per method, ICSD increases per-token actor-update time by 1.7% over the trust-only SDAR allocator. Since actor updates account for roughly one fifth of step time, this corresponds to about 0.4% normalized end-to-end overhead.

Setting	ALFWorld	WebShop	Search-QA
Train / validation batch	16 / 128	16 / 128	128 / 512
Rollouts per task	8	8	8
Maximum interaction steps	50	15	4
Maximum prompt / response length	2048 / 512	4096 / 512	4096 / 512
Actor mini-batch size	256	64	256
Actor learning rate	1×10^{-6}	1×10^{-6}	1×10^{-6}
Ratio clip	0.2	0.2	0.2
KL coefficient	0.01	0.01	0.001
Distillation coefficient λ	0.01	0.01	0.01
ICSD training updates	150	150	200

Table A1: SDAR-aligned protocol used for ICSD training.

Clipping exposure. The influence score is defined on the unclipped surrogate term, while the deployed loss clips the policy ratio. Across the four frozen ALFWorld evidence batches (544,427 valid response tokens), 0.128% of tokens fall outside the clip interval and 0.069% lie where the clipped surrogate is locally flat, so that u_t overstates the local response. These tokens receive 0.103% of the composite coefficient mass and 0.984% of the absolute weighted distillation mass $c_t|\Delta_t|$; no token reaches the negative-advantage dual-clip cap. The flat-token fraction is stable across batches (0.058–0.075%), and the pooled ratio distribution concentrates near one (mean 1.0006, standard deviation 0.0254). These statistics describe the analyzed step-135 batches rather than every training stage; within them, the unclipped approximation misprices only a negligible portion of the auxiliary update.

C.3 Allocator Implementation

Trust and influence calibration. The inherited trust role $g_t = G_{\mathrm{trust}}(\Delta_t)$ is instantiated as a fitted asymmetric location–scale CDF of the sampled gap. Raw influence values inherit the heavy tails and drifting scale of the advantage, policy ratio, and teacher–student gap. Their supported and opposed sides can also have different spreads. The ICSD-specific map $\Phi_{\mathcal{G}}$ therefore uses the same distribution family with a robust location estimate and separate one-sided scales. Each scale is floored away from zero. Fits use valid response tokens from the current calibration group. Undersized or degenerate groups fall back first to global minibatch statistics and then to a fixed monotone map. The statistics are detached and recomputed for every minibatch. We fit the CDF before applying the fallback rule. The remaining coefficients are then rescaled within each action turn so that their composite mass, together with the unchanged fallback coefficients, exactly equals the turn's base trust mass. Algorithm A1 summarizes the update.

D Full Signal-Combination Audit

Figure A1 compares base trust with the deployed ICSD allocation in the four nonzero sign regions. The row labels report the signs of the sampled teacher–student gap Δ_t and the policy advantage $\widehat{A}_t$.

Algorithm A1 ICSD update for one on-policy minibatch

Require: minibatch $\mathcal{B}$ of valid response tokens grouped into action turns; distillation weight λ; scale floor ε

Roll out π_θ; compute $\widehat{A}_t$, ρ_t, and the sampled student log probabilities.
Evaluate the stop-gradient privileged branch π_θ^+ on the same contexts with input z_t; compute Δ_t and the base trust coefficient $g_t = G_{\text{trust}}(\Delta_t)$.
Compute the detached influence $u_t = \widehat{A}_t \rho_t \Delta_t$.
Fit $(\hat{\mu}, \hat{b}_-, \hat{b}_+)$ on the calibration group by the median and one-sided means floored at ε; set $\tilde{m}_t = \Phi_{\mathcal{G}}(u_t)$. Undersized groups fall back to minibatch statistics.
Form $\mathcal{D} = \{t : \widehat{A}_t < 0,\ \Delta_t < 0\}$ and $\mathcal{C} = \mathcal{B} \setminus \mathcal{D}$.
for each action turn q **do**
  if $\sum_{j \in \mathcal{C}_q} g_j \tilde{m}_j > \varepsilon$ **then**
    $\alpha_q \leftarrow \left(\sum_{j \in \mathcal{C}_q} g_j\right) / \left(\sum_{j \in \mathcal{C}_q} g_j \tilde{m}_j\right)$
    $c_t \leftarrow g_t$ on $\mathcal{D}_q$; $c_t \leftarrow \alpha_q\, g_t \tilde{m}_t$ on $\mathcal{C}_q$
  else
    $c_t \leftarrow g_t$ for all $t \in q$ {degenerate turn: retain base trust}
  end if
end for
Detach $\{c_t\}$ and update θ with the allocation objective.

Ensure: $\sum_{t \in q} c_t = \sum_{t \in q} g_t$ for every action turn

Trusted-conflict mass. For nonnegative token weights w_t, define

$$\mathrm{TCM}_w(\mathcal{B}) = \frac{\sum_{t \in \mathcal{B}} w_t\, \mathbf{1}[\Delta_t > 0,\ |u_t| > \varepsilon,\ u_t < 0]}{\sum_{t \in \mathcal{B}} w_t\, \mathbf{1}[\Delta_t > 0,\ |u_t| > \varepsilon]}. \tag{A12}$$

Here ε removes zero-advantage tokens, $w_t = g_t$ for SDAR, and $w_t = c_t$ for ICSD. This normalized share measures allocation rather than raw coefficient scale. Across four matched ALFWorld batches, TCM falls from 60.1% under SDAR to 37.8% under ICSD. The paired reduction is 22.3 points (95% action-turn bootstrap CI $[22.1, 22.6]$) and appears in every batch.

Gradient compatibility. Across sixteen frozen batches, ICSD increases cosine compatibility with the RL gradient by 0.192 on average (95% batch-bootstrap CI $[0.147, 0.240]$), with a positive change in every batch. Permuting the same multipliers across tokens preserves their distribution and coefficient mass but leaves only a 0.020 gain. The pairing-specific difference is 0.172 (95% CI $[0.131, 0.217]$), tying the geometric improvement to token assignment rather than aggregate scale.

Across the frozen batches, ICSD assigns 6.2 percentage points more mass to $\Delta_t > 0, \widehat{A}_t > 0$, 7.5 points less to $\Delta_t > 0, \widehat{A}_t < 0$, and 1.5 points less to $\Delta_t < 0, \widehat{A}_t > 0$. When both signals are negative, the deployed coefficient remains equal to base trust, so the difference is zero rather than a spurious positive shift.

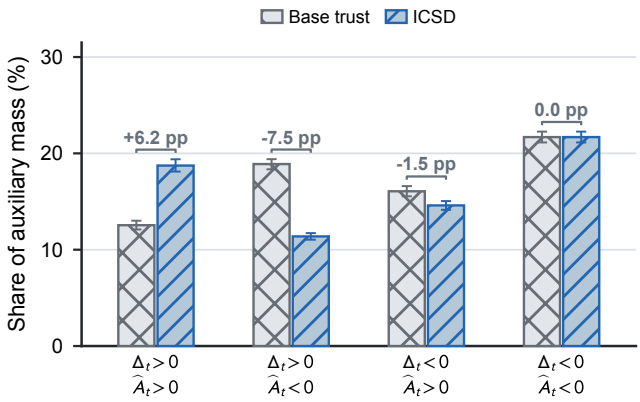

Figure A1: Deployed allocation across the four sign regions. ICSD moves auxiliary mass toward jointly supported tokens and away from sign-conflicting regions, while leaving the double-negative region unchanged. Bars and whiskers show mass shares and 95% action-turn bootstrap intervals over four ALFWorld batches (544,427 tokens); brackets give point differences from base trust.

E Extended Learning Curves

Figure A2 reports the component variants used in the main paper's ablation table.

Figure A3 adds a temporal view to the frozen-batch mechanism audit. The allocation signal is noisy at the minibatch level, especially under GRPO, yet the five-step trends remain below the corresponding SDAR traces for most of training. The difference widens over the final 30 steps under GRPO (46.6% versus 56.1%) and narrows under GiGPO (47.8% versus 49.2%).

F Continuous-Landscape Construction Details

For panels (a–b) of the main paper's mechanism-analysis figure, let c_t denote the deployed coefficient and let q_T and q_R be the empirical mid-ranks of Δ_t and $\widehat{A}_t$ within each environment. For a joint rank bin $\mathcal{B}_{ij}$, the plotted color is the locally smoothed normalized allocation intensity

$$\widehat{I}_{ij} = \frac{|\mathcal{B}_{ij}|^{-1} \sum_{t \in \mathcal{B}_{ij}} c_t}{|\mathcal{B}|^{-1} \sum_{t \in \mathcal{B}} c_t}. \tag{A13}$$

Thus, $\widehat{I}_{ij} = 1$ denotes the environment-average deployed allocation. The gray contours show token density and do not enter the allocator; the rank-transformed axes expose dependence between the two signals rather than their raw scale.

Panel (c) directly measures redistribution relative to the base trust allocation inside the teacher-supported set $\mathcal{S}^+ = \{t : \Delta_t > 0\}$. Negative and positive advantages are ranked separately and divided into five equal-frequency bins on each side of zero; exact zeros remain a separate neutral point and are never jittered into an artificial order. For a bin $\mathcal{K}_k$, the

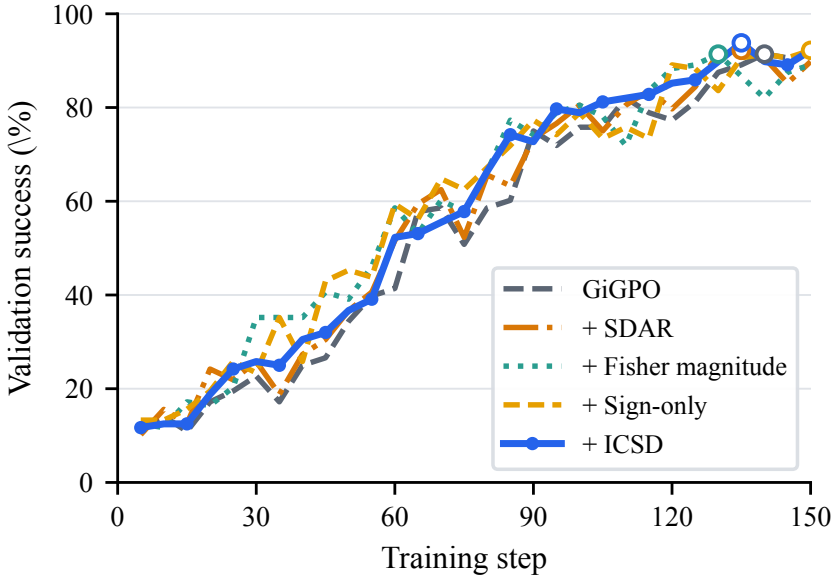

Figure A2: Validation curves for matched ALFWorld allocation variants with Qwen2.5-1.5B, GiGPO, and 150 training updates under an identical training configuration.

vertical axis is

$$\Delta s_k = 100 \left(\frac{\sum_{t \in \mathcal{K}_k} c_t}{\sum_{t \in \mathcal{S}^+} c_t} - \frac{\sum_{t \in \mathcal{K}_k} g_t}{\sum_{t \in \mathcal{S}^+} g_t} \right), \qquad \text{(A14)}$$

in percentage points per bin. Confidence intervals use action-turn bootstrap resampling over the four evidence batches. Since $\mathcal{S}^+$ excludes $\mathcal{D}$, Equation (A14) describes the deployed allocation exactly within this conditional normalization.

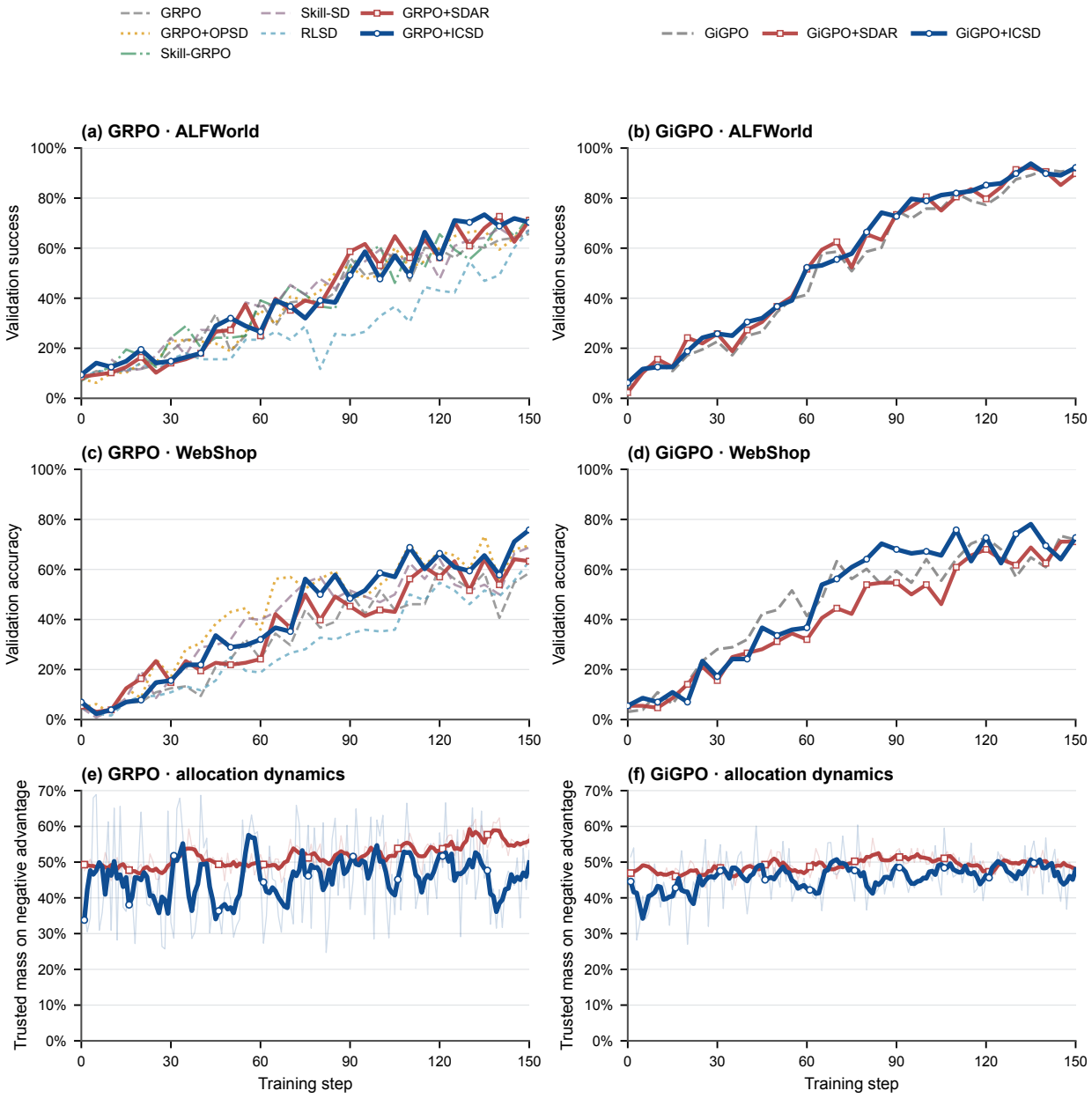

Figure A3: Training dynamics at Qwen2.5-1.5B. Panels (a–b) show ALFWorld, panels (c–d) show WebShop, and panels (e–f) show the share of trusted auxiliary mass assigned to negative-advantage tokens. Pale lines are per-step values and dark lines are centered five-step means. This share describes allocation, not token utility.

		ALFWorld							Search-QA								WebShop	
Family	Method	Pick	Look	Clean	Heat	Cool	Pick2	Avg	NQ	Triv	Pop	Hotp	2Wk	MuS	Bam	Avg	Score	Acc.
Qwen2.5-1.5B-Instruct																		
No RL	Vanilla	11.1	0.0	6.2	0.0	0.0	4.2	5.5	10.5	25.4	17.8	8.3	14.8	2.2	6.0	12.1	17.8	5.5
	Skill-Prompt*	3.4	16.7	12.9	5.3	0.0	5.0	6.2	6.2	18.1	8.7	5.4	10.1	1.5	4.0	7.7	20.8	1.6
	OPSD	26.3	16.7	9.1	6.7	9.1	5.3	14.1	0.0	0.0	0.0	0.1	0.2	0.0	0.0	0.0	22.3	10.2
Hybrid	Skill-SD	88.9	57.1	60.0	80.0	65.0	43.5	68.0	16.7	35.4	14.8	16.7	25.2	3.9	11.2	17.7	80.6	68.8
	RLSD	78.1	41.7	68.2	64.3	77.3	57.7	67.2	13.9	29.4	12.4	14.2	21.1	3.0	9.5	14.8	80.5	63.3
Ours	GiGPO+ICSD	**93.1**	**91.7**	**92.0**	**100.0**	**95.5**	**92.3**	**93.8**	**17.5**	**36.8**	**14.5**	**17.7**	**24.4**	**4.1**	**12.7**	**18.2**	**92.2**	**78.1**
Qwen2.5-3B-Instruct																		
No RL	Vanilla	44.4	11.1	6.2	15.4	28.6	12.5	21.9	24.6	48.1	31.0	26.3	25.3	7.2	59.7	31.7	6.7	0.8
	Skill-Prompt*	51.7	66.7	48.4	0.0	4.3	10.0	28.9	23.7	46.2	30.6	24.4	22.1	7.5	12.5	23.9	0.2	0.8
	OPSD	48.8	41.7	16.7	0.0	15.8	16.7	28.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	11.3	3.1
Skill/RL	Skill-GRPO	88.9	71.4	58.8	70.6	40.7	29.2	60.2	43.5	58.8	43.0	36.8	32.2	11.7	12.5	34.1	77.3	60.9
	Skill-GRPO*	94.3	57.1	100.0	66.7	73.1	57.1	80.5	44.3	59.6	**44.3**	39.0	36.1	14.5	14.9	36.1	76.3	66.4
Hybrid	Skill-SD	88.2	50.0	96.2	52.4	65.0	57.9	73.4	**44.4**	**60.4**	44.0	39.5	**40.4**	15.4	64.9	**44.1**	75.9	64.0
	RLSD	87.9	75.0	90.9	75.0	73.1	68.4	79.7	41.5	58.6	42.3	**40.4**	40.2	**16.8**	**66.9**	43.8	84.4	66.4
Ours	GiGPO+ICSD	**100.0**	**92.9**	**100.0**	**86.7**	**95.0**	**91.3**	**95.3**	43.0	60.3	43.9	38.2	40.0	13.4	64.5	43.3	**88.7**	**75.8**
Qwen2.5-7B-Instruct																		
No RL	Vanilla	36.1	22.2	3.1	0.0	0.0	0.0	12.5	25.2	50.8	29.5	29.0	29.0	10.4	63.7	33.9	5.9	1.6
	Skill-Prompt*	51.7	50.0	32.3	5.3	4.3	0.0	23.4	30.9	52.1	32.7	32.7	27.9	12.7	66.1	36.4	1.7	0.8
	OPSD	50.0	60.0	22.7	21.4	17.6	9.5	32.8	8.8	8.6	17.5	2.5	4.2	0.5	1.2	6.2	4.5	2.3
Skill/RL	Skill-GRPO	88.5	66.7	65.2	61.1	57.7	73.1	69.5	45.2	63.7	45.7	43.1	43.3	19.6	21.4	40.3	80.4	71.9
	Skill-GRPO*	100.0	83.3	**96.4**	83.3	75.0	78.9	88.3	44.8	63.0	45.1	43.7	43.7	20.5	71.4	47.5	87.0	81.2
Hybrid	Skill-SD	93.9	**93.8**	90.9	100.0	69.2	68.4	85.1	47.1	64.5	47.8	44.2	42.1	20.2	69.0	47.8	86.1	76.5
	RLSD	100.0	87.5	92.3	58.8	80.0	65.2	82.0	46.8	63.0	44.4	**45.5**	**48.9**	**21.5**	73.0	49.0	87.4	77.3
Ours	GiGPO+ICSD	**100.0**	91.7	90.9	**100.0**	**95.5**	**96.2**	**96.1**	**47.8**	**64.6**	**48.5**	44.2	45.3	20.1	**73.3**	**49.1**	**93.1**	**84.4**
Qwen3-1.7B-Instruct																		
No RL	Vanilla	25.0	22.2	3.1	0.0	21.4	4.2	12.5	29.4	46.9	37.0	23.5	19.6	6.4	10.5	24.8	46.5	4.7
	Skill-Prompt*	10.3	50.0	16.1	0.0	0.0	5.0	9.4	29.4	46.5	36.2	22.9	20.8	4.3	10.1	24.3	23.0	2.3
	OPSD	26.3	33.3	9.1	0.0	4.5	5.3	14.1	4.2	8.3	4.6	6.6	15.3	0.7	1.2	5.8	47.4	9.3
Skill/RL	Skill-GRPO	27.6	54.5	22.7	27.3	0.0	19.2	21.1	39.2	58.6	43.9	35.2	28.2	11.5	66.1	40.4	73.4	46.1
	Skill-GRPO*	31.4	42.9	51.9	8.3	41.5	7.1	28.1	38.0	58.4	43.9	**36.3**	29.0	12.5	**66.9**	40.7	80.4	50.0
Hybrid	Skill-SD	52.9	37.5	69.2	42.9	60.0	36.8	52.3	39.1	57.5	45.4	34.8	**34.1**	10.7	64.1	40.8	**81.8**	53.9
	RLSD	50.0	37.5	61.5	19.0	50.0	21.1	42.2	38.6	57.3	43.0	34.5	**34.1**	11.5	65.3	40.6	74.0	50.8
Ours	GiGPO+ICSD	**94.1**	**62.5**	**100.0**	**71.4**	**85.0**	**57.9**	**82.8**	**43.6**	**59.7**	**47.0**	**36.3**	31.3	**13.2**	66.1	**42.5**	81.5	**68.0**

Table A2: Full per-task ALFWorld, per-subset Search-QA, and WebShop results corresponding to the aggregate table in the main paper. Starred methods retain retrieved skills during validation.