NATURAL-GRADIENT CALIBRATION OF MULTI-TURN ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation trains a language model by matching a teacher on the student's own rollouts, and the dense token-level signal complements the sparse terminal reward of reinforcement learning. Rollouts of multi-turn agents divide into turns at tool-call boundaries. Recent methods reweight the teacher-matching loss turn by turn from a gap between teacher and student probabilities, from the terminal reward, or from a depth schedule, without estimating how the teacher supervision at a turn contributes to task success. We view on-policy distillation as a policy-gradient step whose advantage is a proxy, the weighted distillation signal, so a turn-level weight scales that proxy toward the true advantage. The weight that fits the proxy best under the student's action distribution is the covariance of the signal with the return divided by the variance of the signal, which is the projection coefficient of the reward gradient onto the distillation direction in the Fisher metric. The projection splits the policy gradient into a calibrated distillation direction and a residual identified by the return alone, so calibrated distillation is the part of reinforcement learning that the teacher supplies. The nonnegative parts of the per-turn projections are also the assignment that gains the most reward in the second-order KL geometry of one update. We instantiate the calibration as ***T**urn-Level **A**dvantage **C**alibration in the **T**rust Region* (TACT). The loss already computes the divergence at each position, and adding it to the token log-ratio makes one rollout an unbiased sample of both the covariance and the variance of every turn it visits. TACT therefore estimates the weights from the training batch, pooled over classes of turns, and adds a reinforcement step on the residual in the classes where the signal explains little of the return. Experiments on ALFWorld, ScienceWorld, WebShop, Search-QA, and τ-bench, on the same rollouts for every method, show that TACT exceeds the strongest existing rule on every environment, by 4.9 points on average over the rule that uses the return, at the same cost as on-policy distillation. The code is available at TACT Codebase.

1 INTRODUCTION

On-policy distillation (OPD) trains a language model to match a teacher on the student's own rollouts. For multi-turn agents, these rollouts are typically divided into turns at tool-call boundaries, with each turn comprising an action block and the resulting environment response. Complementing reinforcement learning from terminal rewards (Wang et al., 2025; Qi et al., 2025; Xi et al., 2025; Li et al., 2026b), OPD provides teacher supervision at every token. By querying the teacher at states visited by the student, on-policy distillation addresses the exposure bias inherent in offline imitation (Bengio et al., 2015; Ranzato et al., 2015; Agarwal et al., 2024; Gu et al., 2024; Lu & Lab, 2025).

Recent work demonstrates the benefits of reweighting the teacher-student matching loss across turns (Zhong et al., 2026; Tan et al., 2026; Zheng & Jiang, 2026), e.g., strategies that decay the weights with turn depth (Xie et al., 2026; Tan et al., 2026; Wang et al., 2026a). Other approaches derive weights from either teacher–student probability discrepancies (Zhong et al., 2026; Xu et al., 2026; Zheng & Jiang, 2026) or terminal-reward supervision (Yang et al., 2026a). However, existing methods assign weights using predefined rules or coarse-grained signals, without explicitly estimating how teacher supervision at each turn contributes to task success. As shown in Figure 2, comparable teacher–student divergence can correspond to positive, negligible, or negative contributions to terminal reward. Moreover, prefix likelihood ratios can decay with turn depth as teacher–student

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.

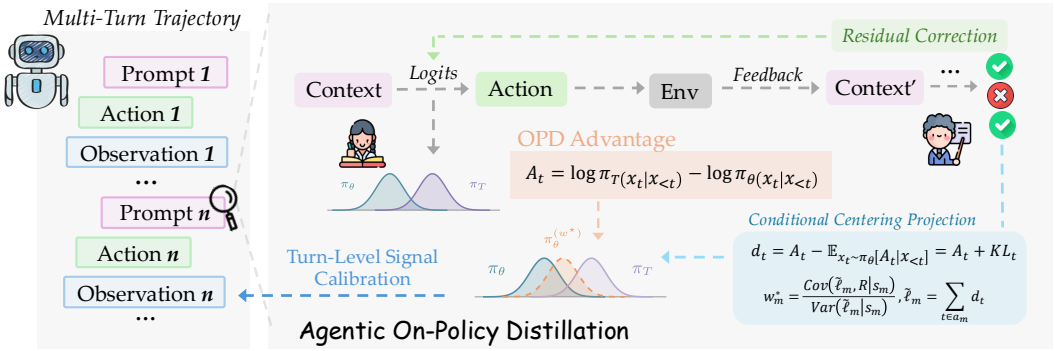

Figure 1: Overview of TACT. The trajectory alternates actions and observations, the teacher scores every action of the student, and the environment returns the outcome at the end. TACT centers the token log-ratio by the divergence at its position, projects the return on the centered signal of each turn, and uses the coefficient as the weight of the turn, while the residual step updates the action with the part of the return that the signal leaves unexplained.

divergence accumulates, even when the value of teacher guidance remains unchanged. These observations motivate a systematic approach that calibrates turn-level teacher supervision according to its individual contribution to terminal reward.

We view on-policy distillation as a policy-gradient step in which a proxy advantage measures the teacher's preference for an action. This proxy advantage must then be scaled by a set of turn-level weights toward the true advantage, which reflects how much that action improves task success. We formulate the choice of these weights as a least-squares problem at each turn. The weight that brings the proxy closest to the true advantage under the distribution of the student's own action blocks is the covariance of the distillation signal with the return divided by the variance of the signal. The inner product behind this fit is the Fisher information metric of the block distribution, so the weight is the projection coefficient of the reward gradient onto the distillation direction, and the variance in its denominator is the Fisher information of the turn. In the same metric the KL divergence of an update separates over turns, so the per-turn projections are also the assignment of one bounded update that gains the most reward.

Our approach naturally splits the policy gradient in on-policy distillation into two: a calibrated distillation direction computed from the teacher-student discrepancy in the cached logits and a reward-driven residual identified from the terminal reward alone. The analysis characterizes one update, turn by turn, as a natural-gradient step in the second-order KL geometry. The bound on its divergence keeps each update within the range where the first-order description of the reward is valid, and unclipped distillation collapses when it leaves that range (Zhao et al., 2026).

We instantiate the calibration as ***T**urn-Level **A**dvantage **C**alibration in the **T**rust Region* (TACT). The analysis reduces the weight of a turn to two conditional expectations at its starting state, the covariance of the signal with the return and its variance, which TACT estimates from the training batch alone, without extra generation, as Figure 1 shows. More precisely, we show that the conditional mean of the teacher–student token log-ratio is the negative KL divergence already computed by the distillation loss. Adding this divergence then leads to a conditionally zero-mean token signal, whose sum over a turn is also zero-mean given the turn's starting state. Multiplying this turn-level signal by its rollout return gives an unbiased estimate of the conditional covariance, and its square an unbiased sample of the variance. Thus, each rollout supplies samples of both quantities for every turn it visits. TACT pools these samples across classes of turns defined by their starting states and uses the estimated covariance-to-variance ratio, clipped to be nonnegative and globally scaled, to weight the distillation loss. TACT adds a reward-driven update on the estimated residual in the classes where the signal leaves substantial reward variation unexplained.

- We formulate turn-level reweighting as the rescaling of a proxy advantage reflecting both the teacher's preference for an action and how much that action improves task success. The formulation projects the reward gradient onto the distillation direction and admits a natural-gradient interpretation.

- We show that the return multiplied by the centered signal identifies the projected gain, while probability gaps and terminal returns alone cannot determine it, and that a local estimate of the variance avoids the depth drift of accumulated ratios.
- We build TACT from single-sample estimators made unbiased by the centered signal, bound the first-order reward of the estimated class weights, and show on five agentic environments that it exceeds the strongest existing rule by 4.9 points on average at the same cost as on-policy distillation.

2 RELATED WORK

On-policy distillation. Sampling from the student and scoring under the teacher removes the exposure bias of sequence-level distillation (Hinton et al., 2015; Kim & Rush, 2016; Ross et al., 2011; Lin et al., 2020; Agarwal et al., 2024), and later work changed the divergence and the sampling schedule (Gu et al., 2024; Wen et al., 2023; Ko et al., 2024; Xu et al., 2025) or the target (Yang et al., 2026b), while keeping the per-position weight unchanged. Privileged teachers, the setting of our experiments, condition on information withheld from the student (Penaloza et al., 2026; Zhao et al., 2026), and anchored residual guidance separates the foresight of such a teacher from a locally compatible anchor (Zhang, 2026b). Our residual is instead the part of the student's own advantage that the teacher does not explain. Gu et al. (2024) and Yang et al. (2026b) write the objective as a policy-gradient step whose advantage is the token log-ratio, and our analysis builds on that form.

Turn-level reweighting for agents. Recent methods reweight the teacher-matching loss per tool call from a gap between teacher and student probabilities (Zhong et al., 2026), gating it by sign consistency (Xu et al., 2026) or by blockwise policy drift (Zheng & Jiang, 2026), and all three compute their statistics from the two policies alone. IW-OPD upweights each token by the accumulated prefix ratio (Xie et al., 2026). ATOD and TCOD anneal turn-level profiles fixed before the rollout (Tan et al., 2026; Wang et al., 2026a), and TCOD represents this family in our comparison. ReOPD replays teacher prefixes under a step-decaying schedule so that the teacher supervises states near its own support (Liao et al., 2026). Verifiable-reward training (Shao et al., 2024; Guo et al., 2025) extends to distillation, where RLSD derives a per-token magnitude from the policy difference and takes one sign per rollout from the terminal reward (Yang et al., 2026a). TACT takes the scale of the proxy advantage from the covariance of the signal with the return, estimated on the batch, while RLSD uses the sign of one return per rollout. Its reinforcement term is the residual of the projection, while SOD sums the two objectives at a fixed ratio.

Credit assignment and trust regions. Step-level credit assignment addresses the same question in agent reinforcement learning, with Monte Carlo value estimates (Kazemnejad et al., 2024) or group-relative advantages at the step level (Feng et al., 2026; Zhang et al., 2026; Wang et al., 2026b; Zhang, 2026a), and GEAR reports that a token-divergence signal is least shared within segments cut at tool-call boundaries (Li et al., 2026a). Bounding the change in return by a divergence underlies conservative policy iteration and trust-region methods (Kakade & Langford, 2002; Schulman et al., 2015; 2017), which constrain one global update per batch. Natural-gradient methods scale the update by the Fisher information (Amari, 1998; Kakade, 2001), and TACT applies both ideas turn by turn.

3 TURN-LEVEL CREDIT ASSIGNMENT IN ON-POLICY DISTILLATION

3.1 ON-POLICY DISTILLATION AS POLICY GRADIENT WITH A PROXY ADVANTAGE

We model a rollout τ on a task input x as a sequence of turns $m = 1, \dots, M$ cut at tool-call boundaries. Turn m starts at the state $s_m = (x, a_1, o_1, \dots, a_{m-1}, o_{m-1})$, the student π_θ with parameters θ draws the action block $a_m \sim \pi_\theta(\cdot \mid s_m)$ token by token, the environment returns the response o_m, and the completed rollout receives the return $R(\tau) \in [0, 1]$, and a policy π attains the expected reward $\eta(\pi) = \mathrm{E}_{\tau\sim\pi}[R(\tau)]$. The teacher π_T is a larger model that reads the context of the student together with a privileged hint withheld from the student, and it places positive probability where the student does; the notation $\pi_T(\cdot \mid c_t)$ leaves the hint implicit.

Write c_t for the prefix of position t of the block, which consists of the state and the earlier tokens of the block, and $\ell_t = \log \pi_T(y_t \mid c_t) - \log \pi_\theta(y_t \mid c_t)$ for the token log-ratio, which Gu et al. (2024)

use as the per-token advantage. The signal of a turn is the token log-ratio centered at its conditional mean,

$$d_t = \ell_t + \mathrm{KL}(\pi_\theta(\cdot \mid c_t) \,\|\, \pi_T(\cdot \mid c_t)), \qquad \tilde{\ell}_m = \sum_{t \in a_m} d_t, \tag{1}$$

since the conditional mean of the log-ratio at a position is the negative divergence there. Each term has zero mean given its prefix, so the block signal has zero mean given the state and its terms are uncorrelated. The implementation clips the token log-ratio so that $|\tilde{\ell}_m| \le L$, and the bounds below use the clipped signal, whose conditional mean moves by at most the expected excess beyond the clip.

The rules that reweight per tool call share a common form, the minimization of a weighted teacher-matching objective,

$$\mathcal{L}(\theta) = \mathrm{E}_{\tau \sim \pi_\theta}\Big[\sum_m w_m \sum_{t \in a_m} \mathrm{KL}(\pi_\theta(\cdot \mid c_t) \,\|\, \pi_T(\cdot \mid c_t)) \Big], \tag{2}$$

where w_m weights the divergence of turn m, summed over its positions. Every implementation of on-policy distillation keeps the visitation distribution and the weights fixed during the update. Under that convention the score identity gives the descent direction of the objective, and the policy-gradient theorem gives the gradient of the reward,

$$\begin{aligned} -\nabla_\theta \mathcal{L}(\theta) &= \mathrm{E}_{\tau \sim \pi_\theta}\Big[\sum_m w_m \sum_{t \in a_m} d_t \, \nabla_\theta \log \pi_\theta(y_t \mid c_t) \Big], \\ \nabla_\theta \eta(\pi_\theta) &= \mathrm{E}_{\tau \sim \pi_\theta}\Big[\sum_m \sum_{t \in a_m} A^{\pi_\theta}(c_t, y_t) \, \nabla_\theta \log \pi_\theta(y_t \mid c_t) \Big], \end{aligned} \tag{3}$$

with $Q^{\pi_\theta}(c_t, y_t) = \mathrm{E}[R \mid c_t, y_t]$ the expected reward after the token when the student continues and $A^{\pi_\theta}(c_t, y_t) = Q^{\pi_\theta}(c_t, y_t) - \mathrm{E}_{\pi_\theta}[Q^{\pi_\theta}(c_t, y) \mid c_t]$ the student's advantage at the position. The centering leaves the direction unchanged, since the divergence at a position depends on the prefix alone and is therefore a baseline of the score (Williams, 1992). The two gradients differ only in the advantage: on-policy distillation is a policy-gradient step whose advantage at a turn is the weighted signal in place of the true advantage (Gu et al., 2024; Yang et al., 2026b). Every turn-level rule is therefore a rescaling of the proxy advantage, with a scale that existing rules take from the signal alone.

3.2 Calibration as Advantage Projection

The scale that makes the proxy closest to the true advantage over the positions of a turn is a least-squares fit under the student's distribution, with the inner product $\langle f, h \rangle = \mathrm{E}_{\pi_\theta}[\sum_{t \in a_m} f_t h_t \mid s_m]$. That inner product is the Fisher metric of the block distribution at s_m.

Lemma 1 (Calibration as projection)**.** *At state s_m, the scale w that minimizes $\mathrm{E}_{\pi_\theta}[\sum_t (A^{\pi_\theta}(c_t, y_t) - w\, d_t)^2 \mid s_m]$ over the positions t of the block is*

$$w_m^\star = \frac{g_m}{V_m}, \qquad g_m = \mathrm{E}_{\pi_\theta}\big[\tilde{\ell}_m \, R \,\big|\, s_m\big], \qquad V_m = \mathrm{E}_{\pi_\theta}\big[\tilde{\ell}_m^2 \,\big|\, s_m\big], \tag{4}$$

with $V_m \le L^2$ and $|g_m| \le \frac{1}{2}\sqrt{V_m}$. In the Fisher metric at s_m, $\langle \tilde{\ell}_m, A^{\pi_\theta} \rangle = g_m$ and $\|\tilde{\ell}_m\|^2 = V_m$, so $w_m^\star$ is the projection coefficient of the reward gradient onto the distillation direction.

Appendix B proves the lemma and the decomposition that follows it. We call the three quantities the projected gain, the signal variance, and the calibrated weight of the turn. The gain is the covariance of the signal with the return, since the signal has zero conditional mean. For a binary reward it is the success probability at the state times the mean signal of the successful rollouts through it, positive where the blocks the teacher prefers succeed more often.

The projection leaves a remainder: the part of the advantage that the signal does not explain, $r_t = A^{\pi_\theta}(c_t, y_t) - w_m^\star d_t$, is orthogonal to the signal in the same metric. Substituting the decomposition into the policy gradient splits it into two terms,

$$\begin{aligned} \nabla_\theta \eta(\pi_\theta) = &- \mathrm{E}_{\tau \sim \pi_\theta}\Big[\sum_m w_m^\star \, \nabla_\theta \sum_{t \in a_m} \mathrm{KL}(\pi_\theta(\cdot \mid c_t) \,\|\, \pi_T(\cdot \mid c_t)) \Big] \\ &+ \mathrm{E}_{\tau \sim \pi_\theta}\Big[\sum_m \sum_{t \in a_m} r_t \, \nabla_\theta \log \pi_\theta(y_t \mid c_t) \Big], \end{aligned} \tag{5}$$

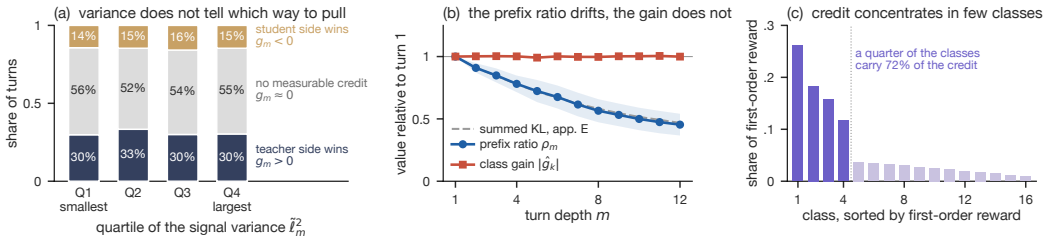

Figure 2: Panel (a) splits each quartile of the signal variance by the sign of the class gain: the turns with positive gain make up 30 to 33% of every quartile. Raising the weight with the variance therefore pulls toward the teacher where it is wrong as often as where it is right. Panel (b): by the twelfth turn the prefix ratio falls to less than half of its first-turn value while the gain magnitude stays at its first-turn level, so a depth-decaying weight discounts late turns for a reason unrelated to their credit. Panel (c): about a quarter of the classes carry 72% of the calibrated first-order reward, led by the classes after a failed call and the latest calls.

the descent direction of the objective at the calibrated weights, evaluated over the whole vocabulary from the cached logits, and a policy gradient whose advantage is the residual. Calibrated on-policy distillation is therefore the part of the policy gradient that lies along the teacher, and the residual is the part that the teacher cannot supply and only the return identifies. The calibrated weight also leaves the residual with the least variance, so the signal is the control variate of the return (Glasserman, 2003).

The gain is a covariance with the return, so it depends on the responses of the environment and not only on the two policies. Replacing the return by one minus the return flips its sign and changes neither the signal nor the student's block distribution, so weights computed from the two policies do not determine the sign of the credit. The terminal return alone gives every turn of a rollout the same sign. Averaging the return multiplied by the centered signal over rollouts identifies the sign, and the next section estimates the gain this way. Rules that accumulate the signal along the sampled prefix inherit a decay with depth at the rate of policy separation, independent of the gain of late turns, as the appendix derives.

3.3 Optimality of the Projection in the Trust Region

The per-turn projections of the lemma are the right joint assignment for one update only if the update separates across turns, since every quantity above is conditional on the state and rescaling an early turn changes the states that later turns visit. The natural-gradient step of the objective (Kakade, 2001) at a turn with weight w_m and step size α moves the block distribution to the exponential reweighting of the student by its signal,

$$\pi_\theta^{(\alpha w_m)}(a \mid s_m) = \frac{\pi_\theta(a \mid s_m)\, e^{\alpha w_m \tilde{\ell}_m(a)}}{\mathrm{E}_{\pi_\theta}\left[e^{\alpha w_m \tilde{\ell}_m} \,\middle|\, s_m\right]}, \tag{6}$$

which is the closed form of a KL-constrained step whose reward is the proxy advantage (Yang et al., 2026b; Xie et al., 2026). As an exponential family in the weight, this distribution has the signal variance as its Fisher information at the student (Amari, 2016). The analysis proceeds at the level of block distributions, which a parameter step realizes to first order (Schulman et al., 2015). We study the reweighted policy $\pi_\theta^{(\alpha \boldsymbol{w})}$ that applies the reweighting at every turn with a weight $w_m(s_m)$ that may depend on the state, and both the divergence and the reward of this update separate over turns.

Proposition 1 (Separation over turns)**.** *For every profile with $\alpha|w_m(s_m)|L \le \frac{1}{2}$, the divergence over rollouts and the reward obey*

$$\begin{aligned} \mathrm{KL}\left(\pi_\theta^{(\alpha \boldsymbol{w})} \,\middle\|\, \pi_\theta\right) &= \tfrac{1}{2}\alpha^2\, \mathrm{E}_{\tau\sim\pi_\theta}\left[\sum_m w_m(s_m)^2\, V_m(s_m)\right] + O(\alpha^3), \\ \eta\left(\pi_\theta^{(\alpha \boldsymbol{w})}\right) - \eta(\pi_\theta) &= \alpha\, \mathrm{E}_{\tau\sim\pi_\theta}\left[\sum_m w_m(s_m)\, g_m(s_m)\right] + O(\alpha^2), \end{aligned} \tag{7}$$

where the cross terms between turns vanish in the first identity because the Fisher information in the coordinates $\alpha \boldsymbol{w}$ is diagonal.

The divergence identity follows from the tower property, since the signals of two turns are uncorrelated, and the reward identity follows from the performance difference identity of Kakade & Langford (2002) with a coupling of the two rollouts. The second-order term contains the effect of the moved states and grows with the number of turns, so the step must shrink as the number of turns grows. We take the total divergence as the trust region of one update.

Proposition 2 (Optimal assignment in the trust region)**.** *For every radius $\Delta > 0$, among the profiles $\boldsymbol{w} \geq 0$ whose second-order divergence does not exceed Δ, the first-order reward is maximized at $w_m(s_m) \propto \max\{w^\star_m(s_m), 0\}$, with one constant for all turns set by the radius whenever some gain is positive.*

At the calibrated weight a turn gains g_m^2/V_m in first-order reward per unit step size and costs half of that in divergence, so the update puts most of its weight on the turns whose signal explains the return.

The update uses $w_m = c \max\{w^\star_m, 0\}$ with a global scale c that absorbs the step size and the radius, so from here on the weight includes the step size and the smallness condition becomes $w_m L \leq \frac{1}{2}$. The turns where the signal is uncorrelated with the return and those where the blocks the teacher prefers score worse receive weight zero. We call the divergence between the estimated and the calibrated update, $\frac{1}{2}c^2 \sum_m (\hat{w}_m - w^\star_m)^2 V_m$, the miscalibration, and by Cauchy–Schwarz it bounds the first-order reward that the estimate loses.

4 TACT: PROJECTION FROM THE BATCH

The gain and the variance are expectations of quantities that each rollout already contains. TACT estimates them from the training batch, pools the estimates over classes of turns that share a weight, uses the ratio as the loss weight, and adds the residual step where the signal explains little of the return, as Algorithm 1 lists.

4.1 SINGLE-SAMPLE ESTIMATES AND CLASSES OF TURNS

On-policy distillation already runs one teacher pass per rollout to score the sampled tokens, and the signal reuses that pass without teacher generation. Both quantities of the lemma are conditional expectations of a function of one turn, so the rollout that visits a state supplies one unbiased sample of each: the signal times the return for the gain, and the squared signal for the variance. Subtracting from the return a baseline that depends on the state alone leaves the expectation unchanged, since the signal has zero conditional mean, and we use the mean return of the batch in each environment as the baseline.

Since each state appears in one rollout, we pool the estimates over classes of turns that share a calibrated weight. The class $k(m)$ of a turn is a function of its state: the index of its tool call, the exit status of the previous call, and the verifier output on the previous block where the environment provides one. Each environment has at most thirteen classes, and the first turn of a rollout forms a class of its own. The class uses the status of the previous call, since the block itself determines the status of the current call and a class defined by it would break the centering. Over the N rollouts of an estimation window the class estimates are

$$\hat{g}_k = \frac{1}{N} \sum_{i=1}^{N} \sum_{\substack{m \in \tau_i \\ k(m)=k}} \tilde{\ell}_m \, (R_i - b_m), \qquad \hat{V}_k = \frac{1}{N} \sum_{i=1}^{N} \sum_{\substack{m \in \tau_i \\ k(m)=k}} \tilde{\ell}_m^2, \qquad \hat{w}_k = \frac{\hat{g}_k}{\hat{V}_k}, \tag{8}$$

where $\hat{w}_k$ is the least-squares slope of the centered return on the signal through the origin over the turns of the class. The estimates $\hat{g}_k$ and $\hat{V}_k$ are unbiased for the class totals $\bar{g}_k$ and $\bar{V}_k$ of the gain and the variance and are free of importance weights. The window is an exponential average over batches with rate β, so the weights follow the student as it moves, and each estimate concentrates around its total at the rate $1/\sqrt{N}$.

We estimate the residual of a turn by $\hat{r}_m = (R - b_m) - \hat{w}_{k(m)} \tilde{\ell}_m$, which is unbiased for the residual advantage at every token of the turn. The share of the return that the signal explains in a class is $\kappa_k = \hat{g}_k^2 / (\hat{V}_k \hat{\sigma}_k^2)$, with $\hat{\sigma}_k^2$ the variance of the centered return over the class, and the residual carries the fraction $1 - \kappa_k$ of that variance.

Algorithm 1 TACT: One Update from the Batch

Input: student π_θ, teacher π_T, classes $k(\cdot)$, window rate β, floor ϵ, scale c, residual gate $\kappa_{\min}$
Output: updated parameters θ

■ Signals from the cached logits

for each rollout $\tau_i \sim \pi_\theta$ of the batch and each turn m **do**
  $\tilde{\ell}_m \leftarrow \sum_{t \in a_m} (\ell_t + \mathrm{KL}_t)$, $k \leftarrow k(m)$ // centered signal
  $G_k \mathrel{+}= \tilde{\ell}_m (R_i - b_m)$, $S_k \mathrel{+}= \tilde{\ell}_m^2$, $Q_k \mathrel{+}= (R_i - b_m)^2$
end for

■ Projection within classes

for each class k **do**
  $(\hat{g}_k, \hat{V}_k, \hat{\sigma}_k^2) \leftarrow \beta\,(\hat{g}_k, \hat{V}_k, \hat{\sigma}_k^2) + (1-\beta)\,(G_k, S_k, Q_k)/N$
  $\hat{w}_k \leftarrow \hat{g}_k / \hat{V}_k$ if $\hat{V}_k \geq \epsilon$, else 0; $\kappa_k \leftarrow \hat{g}_k^2 / (\hat{V}_k \hat{\sigma}_k^2)$ // floor
end for

■ One step in two directions

$w_m \leftarrow c \min\{\max\{\hat{w}_{k(m)}, 0\},\ 1/(2cL)\}$ // calibrated weight
$\hat{r}_m \leftarrow (R_i - b_m) - \hat{w}_{k(m)} \tilde{\ell}_m$ if $\kappa_{k(m)} < \kappa_{\min}$, else 0
$\theta \leftarrow$ descent step on $\mathcal{L}(\theta) - \sum_m \hat{r}_m \log \pi_\theta(a_m \mid s_m)$

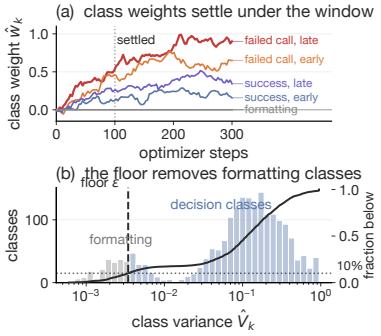

Figure 3: Behavior of Algorithm 1: the class weights settle within a hundred updates under the window, with the classes after a failed call on top, and the floor removes the formatting classes.

The argument of Proposition 2, applied class by class, shows that the ratio of the class totals maximizes the first-order reward within the trust region among the weight profiles that are constant on each class. Finer classes trade sample size for resolution, and the ablations measure this trade.

4.2 Variance Floor and the Reward Guarantee

The update applies the clipped nonnegative part of the class estimate as the weight. The error of the estimated weight has two parts, and both divide by the class variance and grow without bound as the variance approaches zero. The floor ϵ on the class variance therefore bounds every estimated weight, and the appendix bounds the first-order reward lost per excluded class by a term of order $\sqrt{\epsilon}$. We set the floor at the tenth percentile of the class variances of the batch, which excludes the classes of formatting turns, as Figure 3 shows.

Proposition 3 in the appendix bounds the first-order reward of the executed classes, with N taken as the effective sample size under the exponential average. With K classes and the set $\mathcal{K}$ of executed classes, this first-order reward is at least the calibrated first-order reward $c \sum_k \bar{g}_k^2 / \bar{V}_k$ minus an estimation error of order $|\mathcal{K}|\, M^3 L^4 \epsilon^{-2} \sqrt{\log(K/\delta)/N}$ with probability $1 - \delta$ over the rollouts of the window. The error falls with the rollouts and grows with the number of classes.

TACT adds the residual step to the loss on the classes with $\kappa_k < \kappa_{\min}$, a reinforcement step whose advantage is the part of the return that the signal does not explain. The step is off in the classes where the signal explains the return. On a class with negative estimated gain the matching term is off and only the residual step acts, so the update reduces to reinforcement learning with the batch baseline where the teacher is uninformative. Setting the weight to a constant and removing the residual step recovers standard OPD, and TACT has the same cost.

5 Experiments

5.1 Setup

The teacher is Qwen3-32B, which reads the context of the student together with a privileged hint from the task annotation that Table 7 lists per environment. The students are Qwen3-4B and Qwen3-8B in non-thinking mode (Yang et al., 2024; 2025). They share the vocabulary of the teacher, train on the standard split of each environment, and act without the hint at evaluation.

ALFWorld (Shridhar et al., 2020), ScienceWorld (Wang et al., 2022), WebShop (Yao et al., 2022), Search-QA in the retrieval setting of Jin et al. (2025), and τ-bench (Yao et al., 2024) span embodied games, science experiments, web shopping, multi-hop retrieval, and user-facing tool APIs.

Table 1: Task performance on ALFWorld, Search-QA, and WebShop under one matched training budget, the same rollouts for every method. ALFWorld reports success per task type with the count-weighted average, Search-QA exact match per dataset, with NQ, Triv., and Pop single-hop and Hotp., 2Wk, MuS., and Bam. multi-hop, and WebShop the reward score with exact success. Bold and underline mark the best and second best per block and column. ScienceWorld and τ-bench follow in Table 2, and the appendix repeats both tables for the Llama pair.

Method	ALFWorld							Search-QA								WebShop	
	Pick	Look	Clean	Heat	Cool	Pick2	Avg	NQ	Triv.	Pop	Hotp.	2Wk	MuS.	Bam.	Avg	Score	SR
Qwen3-32B (Teacher)	100.0	95.6	100.0	97.4	97.1	96.5	98.1	71.3	83.2	67.6	71.7	74.2	58.7	72.2	71.3	95.1	89.2
Qwen3-8B (Student)																	
Zero-Shot	77.5	66.7	59.4	51.3	50.5	44.7	59.0	34.6	55.9	34.9	30.2	33.9	9.7	36.5	33.7	42.6	18.4
GRPO (Shao et al., 2024)	91.7	67.8	89.7	76.5	74.3	67.1	79.6	43.0	65.4	43.4	40.6	45.1	17.5	40.3	42.2	81.3	68.6
Vanilla OPD (Agarwal et al., 2024)	93.3	73.3	92.3	82.6	81.0	72.9	84.0	45.1	66.8	45.5	43.1	47.3	19.7	42.9	44.3	84.0	72.3
SOD (Zhong et al., 2026)	93.3	74.4	94.2	83.5	81.9	74.1	85.1	46.2	67.8	46.7	45.4	49.5	22.1	45.3	46.1	85.2	73.5
IW-OPD (Xie et al., 2026)	93.3	71.1	92.9	81.7	79.0	71.8	83.3	45.9	67.4	46.2	42.3	46.4	18.6	41.8	44.1	83.5	71.4
TCOD (Wang et al., 2026a)	94.2	75.6	92.9	84.3	82.9	75.3	85.5	45.4	66.6	45.7	42.8	47.4	19.5	43.0	44.3	84.2	72.6
RLSD (Yang et al., 2026a)	94.2	74.4	93.5	83.5	81.9	74.1	85.1	48.6	69.9	49.1	46.9	50.9	23.4	46.7	47.9	86.9	73.7
TACT	**97.5**	**82.2**	**96.8**	**90.4**	**88.6**	**82.4**	**90.7**	**49.4**	**70.3**	**49.7**	**48.5**	**52.6**	**25.2**	**50.7**	**49.5**	**88.9**	**79.2**
Qwen3-4B (Student)																	
Zero-Shot	68.3	58.9	52.9	45.2	44.8	40.0	52.2	29.9	52.7	30.4	25.1	29.1	3.2	31.8	28.9	35.8	13.1
GRPO (Shao et al., 2024)	89.2	60.0	87.1	71.3	68.6	58.8	74.6	38.6	62.7	39.0	36.0	40.8	11.1	35.8	37.7	77.7	62.6
Vanilla OPD (Agarwal et al., 2024)	91.7	66.7	91.6	77.4	76.2	65.9	80.1	41.1	64.4	41.2	38.9	43.4	13.8	38.7	40.2	80.9	66.9
SOD (Zhong et al., 2026)	92.5	68.9	91.6	80.0	78.1	68.2	81.6	42.6	65.7	42.8	42.0	46.3	17.0	41.9	42.6	82.5	69.0
IW-OPD (Xie et al., 2026)	92.5	64.4	91.0	77.4	75.2	64.7	79.6	42.2	65.2	42.8	38.4	42.8	13.1	37.8	40.3	80.6	66.3
TCOD (Wang et al., 2026a)	92.5	68.9	92.3	80.0	79.0	69.4	82.1	41.5	64.3	41.5	38.7	43.6	13.7	38.9	40.3	81.3	67.4
RLSD (Yang et al., 2026a)	92.5	68.9	92.9	80.0	78.1	68.2	81.9	45.5	68.1	**45.8**	43.7	48.0	18.8	43.5	44.8	84.7	69.3
TACT	**95.8**	**77.8**	**94.2**	**87.8**	**85.7**	**76.5**	**87.6**	**45.7**	**68.2**	45.7	**45.2**	**49.9**	**20.6**	**47.0**	**46.0**	**86.4**	**74.5**

Training follows the agentic protocol of Feng et al. (2026), and every method trains on the same rollouts and tool calls. The student samples rollouts at temperature 1.0 with 128 tasks per batch, and training runs 500 optimizer steps per environment. The estimation window averages the class sums over batches at rate 0.9, the residual step is on in the classes whose explained share is below one half, and Table 6 lists the remaining settings. Every reported number is a mean over 5 seeds, every baseline runs in the same protocol from released code or the published weighting rule, and Table 8 counts the generation and teacher-scoring cost of each method per update.

5.2 Main Results

TACT exceeds every existing rule on every environment at both scales on the same rollouts. RLSD, the rule that uses the return, is the closest on Search-QA, WebShop, and τ-bench and trails by 4.9 points on average over the primary metrics, with WebShop measured by success rate. The margin over the strongest rule is largest on τ-bench in Table 2 and smallest on Search-QA, whose four-turn episodes leave three classes to separate the turns. On the same metrics, TACT closes between 41% and 86% of the zero-shot-to-teacher gap at 8B and between 39% and 81% at 4B. Among the existing rules, TCOD leads on ALFWorld and ScienceWorld and stays level with vanilla OPD on Search-QA, and SOD gains the most on Search-QA and τ-bench among the rules that ignore the return. IW-OPD falls below vanilla OPD on four of the five environments at 8B and on three at 4B, as the accumulated ratio discounts the late turns.

5.3 Ablations and Further Analysis

Table 3 removes one ingredient at a time. The two error terms in panel (a) of Figure 4 bound the resulting miscalibration by source. The drop in success follows the measured miscalibration at a correlation of 0.80, and the variant without the floor lies above the trend because the formatting classes then receive weights set by a near-zero denominator, an instability the miscalibration does not record.

Removing the residual step leaves the miscalibration unchanged and costs 1.6 points on Search-QA and 0.7 on ALFWorld. The difference follows the share of the return that the signal explains, which averages 28% over the classes of Search-QA and 45% over those of ALFWorld. The share is lowest on the first call of a rollout and highest after a failed call, as Figure 5 maps. The fixed-coefficient

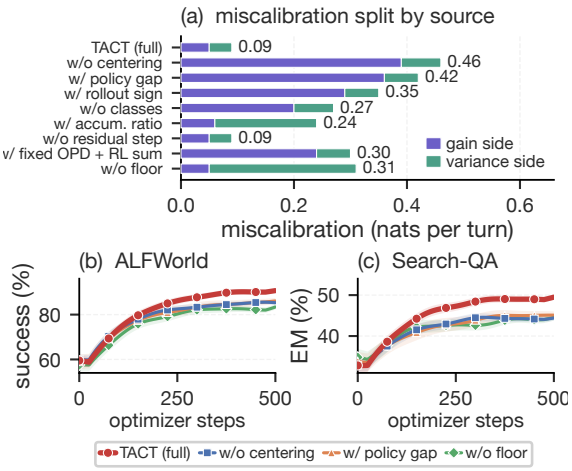

Figure 4: Ablations of Table 3 by error source and by training step with the Qwen3-8B student. Panel (a) shows the gain-side and variance-side error terms, which bound the measured miscalibration of every variant. Panels (b) and (c) show training under the three variants with the largest drop, mean and one standard deviation over 5 seeds.

Table 2: ScienceWorld and τ-bench under the protocol, marks, and method citations of Table 1. ScienceWorld reports the graded score with exact success, and τ-bench reports pass^1 and pass^4 over repeated trials.

Method	ScienceWorld		τ-bench	
	Score	SR	pass^1	pass^4
Qwen3-32B (Teacher)	91.9	89.6	63.2	46.5
Qwen3-8B (Student)				
Zero-Shot	39.2	12.4	22.4	13.0
GRPO	64.4	49.3	26.5	17.5
Vanilla OPD	67.3	53.1	30.6	21.1
SOD	68.2	54.3	32.8	23.0
IW-OPD	66.6	52.2	30.7	21.2
TCOD	68.5	54.6	31.6	21.9
RLSD	68.4	54.5	32.9	23.4
TACT	**74.1**	**61.8**	**39.0**	**25.7**
Qwen3-4B (Student)				
Zero-Shot	35.8	7.9	14.6	8.1
GRPO	59.0	42.4	19.5	12.9
Vanilla OPD	62.6	47.0	24.3	16.8
SOD	64.0	48.8	27.1	19.1
IW-OPD	62.0	46.3	24.7	17.0
TCOD	64.2	49.0	25.7	17.9
RLSD	63.7	48.5	27.6	19.6
TACT	**70.0**	**56.6**	**33.6**	**22.2**

sum of OPD and a reinforcement term used by SOD recovers part of that difference at a larger variance of the update, since the full return enters the update in place of the residual.

Among the replacements of the gain, removing the centering costs the most: the covariance of the raw log-ratio with the return mixes the credit of a turn with the difficulty of the states its class visits. The policy-gap variant makes the weight a function of the variance alone and discards the sign of the credit, and the rollout-sign variant gives every turn of a rollout the same direction. Collapsing to one class per environment keeps the calibration and loses the resolution across turns, which costs three points on ALFWorld. The variance-side replacements fail more mildly: the accumulated ratio inherits the depth decay of the prefix ratio and underweights late turns.

Table 3: Ablations with the Qwen3-8B student under the main protocol. Miscal. is the miscalibration against the class weights of the whole run, in nats per turn.

Variant	ALFWorld ↑	Search-QA ↑	Miscal. ↓
TACT (full)	**90.7**	**49.5**	**0.09**
w/o centering	85.3	44.4	0.46
w/ policy gap	86.1	45.1	0.42
w/ rollout sign	86.8	45.8	0.35
w/o classes	87.7	46.7	0.27
w/ accum. ratio	88.1	47.3	0.24
w/o floor	83.4	44.6	0.31
w/o residual step	90.0	47.9	**0.09**
w/ fixed OPD + RL sum	88.3	47.2	0.30

6 Conclusion

We calibrated the turn-level weight of on-policy distillation as the projection of the reward gradient onto the distillation direction, with the residual of that projection as the reinforcement term the teacher cannot supply. TACT estimates both from the batch with a centered signal and exceeds the strongest existing rule on five environments by 4.9 points on average, at the same cost as on-policy distillation. The ablations trace the gain to the centered covariance in the weight, which neither the policy gap nor the rollout sign replaces, and show that the residual step matters most where the signal explains the least of the return. Calibrated distillation is therefore the part of reinforcement learning that the teacher supplies, and the calibration sets per class how much of the update comes from the teacher and how much from the return alone. The analysis is first order in the reward and second order in the divergence, and extending it to the moved states beyond second order is the next step.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing, LaTeX formatting, and the drafting of figure-generation scripts. The authors reviewed and revised every output of these tools and take full responsibility for all content of the paper.

ETHICS STATEMENT

This work studies the distillation of tool-use agents on publicly available environments and benchmarks. It involves no human subjects, personal data, or sensitive content, and we identify no concern under the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

Section 5.1 specifies the models, the environments, the baselines, and the evaluation protocol. Appendix I lists the training and calibration settings, the environment records, and the class definition of a turn, and the preceding appendices give the derivations behind the stated results. Code and configurations are available at `https://anonymous.4open.science/r/TACT_ICLR-BD0D/README.md`.

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*, volume 2024, pp. 21246–21263, 2024.

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

Shun-ichi Amari. *Information geometry and its applications*. Springer, 2016.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in neural information processing systems*, 28, 2015.

J Frédéric Bonnans, J Charles Gilbert, Claude Lemaréchal, and Claudia A Sagastizábal. *Numerical optimization: theoretical and practical aspects*. Springer, 2006.

Víctor Elvira, Luca Martino, and Christian P Robert. Rethinking the effective sample size. *International Statistical Review*, 90(3):525–550, 2022.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for llm agent training. *Advances in Neural Information Processing Systems*, 38:46375–46408, 2026.

Paul Glasserman. Monte carlo methods in financial engineering. 2003. URL `https://api.semanticscholar.org/CorpusID:60896016`.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the nineteenth international conference on machine learning*, pp. 267–274, 2002.

Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. Vineppo: Refining credit assignment in rl training of llms. *arXiv preprint arXiv:2410.01679*, 2024.

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.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models. *arXiv preprint arXiv:2402.03898*, 2024.

Augustine Kong. A note on importance sampling using standardized weights. *University of Chicago, Dept. of Statistics, Tech. Rep*, 348:14, 1992.

Sijia Li, Yuchen Huang, Zifan Liu, Yanping Li, Jingjing Fu, Li Zhao, Jiang Bian, Ling Zhang, Jun Zhang, and Rui Wang. Gear: Granularity-adaptive advantage reweighting for llm agents via self-distillation. *arXiv preprint arXiv:2605.11853*, 2026a.

Yu Li, Sizhe Tang, and Tian Lan. Reason in chains, learn in trees: Self-rectification and grafting for multi-turn agent policy optimization. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 4673–4686, 2026b.

Baohao Liao, Hanze Dong, Christof Monz, Xinxing Xu, Li Dong, and Furu Wei. Multi-turn on-policy distillation with prefix replay. *arXiv preprint arXiv:2607.04763*, 2026.

Alexander Lin, Jeremy Wohlwend, Howard Chen, and Tao Lei. Autoregressive knowledge distillation through imitation learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6121–6133, 2020.

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

Emiliano Penaloza, Dheeraj Vattikonda, Nicolas Gontier, Alexandre Lacoste, Laurent Charlin, and Massimo Caccia. Privileged information distillation for language models. *arXiv preprint arXiv:2602.04942*, 2026.

Zehan Qi, Xiao Liu, Iat Long Iong, Hanyu Lai, Xueqiao Sun, Jiadai Sun, Xinyue Yang, Yu Yang, Shuntian Yao, Wei Xu, et al. Webrl: Training llm web agents via self-evolving online curriculum reinforcement learning. In *International Conference on Learning Representations*, volume 2025, pp. 79791–79821, 2025.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*, 2015.

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. JMLR Workshop and Conference Proceedings, 2011.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. Pmlr, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*, 2020.

Qitai Tan, Zefang Zong, Mo Li, Yipeng Shi, Yang Li, and Peng Chen. Atod: Annealed turn-aware on-policy distillation for multi-turn agentic tasks. 2026. URL https://api.semanticscholar.org/CorpusID:289669554.

Jiaqi Wang, Wenhao Zhang, Weijie Shi, Yaliang Li, and James Cheng. Tcod: Exploring temporal curriculum in on-policy distillation for multi-turn autonomous agents. *arXiv preprint arXiv:2604.24005*, 2026a.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. Scienceworld: Is your agent smarter than a 5th grader? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11279–11298, 2022.

Yunan Wang, Minghui Song, Zihan Zhang, Shaohan Huang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. Group-graph policy optimization for long-horizon agentic reinforcement learning. *arXiv preprint arXiv:2606.22995*, 2026b.

Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Xing Jin, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, et al. Ragen: Understanding self-evolution in llm agents via multi-turn reinforcement learning. *arXiv preprint arXiv:2504.20073*, 2025.

Yuqiao Wen, Zichao Li, Wenyu Du, and Lili Mou. F-divergence minimization for sequence-level knowledge distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10817–10834, 2023.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Zhiheng Xi, Jixuan Huang, Chenyang Liao, Baodai Huang, Honglin Guo, Jiaqi Liu, Rui Zheng, Junjie Ye, Jiazheng Zhang, Wenxiang Chen, et al. Agentgym-rl: Training llm agents for long-horizon decision making through multi-turn reinforcement learning. *arXiv preprint arXiv:2509.08755*, 2025.

Yan Xie, Sijie Zhu, Tiansheng Wen, Bo Chen, and Yifei Wang. On the position bias of on-policy distillation. *arXiv preprint arXiv:2606.22600*, 2026.

Haoran Xu, Hongyu Wang, Yifei Gao, Jiaze Li, Xiaofeng Zhang, and Xiaosong Yuan. Sg-opd: Sign-gated on-policy distillation via sign-consistency gating and phased teacher sampling. *ArXiv*, abs/2606.09304, 2026. URL https://api.semanticscholar.org/CorpusID:289099145.

Wenda Xu, Rujun Han, Zifeng Wang, Long Le, Dhruv Madeka, Lei Li, William Wang, Rishabh Agarwal, Chen-Yu Lee, and Tomas Pfister. Speculative knowledge distillation: Bridging the teacher-student gap through interleaved sampling. In *International Conference on Learning Representations*, volume 2025, pp. 64616–64646, 2025.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr. *arXiv preprint arXiv:2604.03128*, 2026a.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation. *arXiv preprint arXiv:2602.12125*, 2026b.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35:20744–20757, 2022.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.

Chenchen Zhang. From reasoning to agentic: Credit assignment in reinforcement learning for large language models. *arXiv preprint arXiv:2604.09459*, 2026a.

Wenhao Zhang. Beyond absolute imitation: Anchored residual guidance for privileged on-policy distillation. *arXiv preprint arXiv:2606.10385*, 2026b.

Zijing Zhang, Ziyang Chen, Mingxiao Li, Zhaopeng Tu, and Xiaolong Li. Rlvmr: Reinforcement learning with verifiable meta-reasoning rewards for robust long-horizon agents. In *International Conference on Learning Representations*, volume 2026, pp. 114676–114700, 2026.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

Liwen Zheng and Haiyun Jiang. Blockwise policy-drift gating for on-policy distillation. *arXiv preprint arXiv:2606.24084*, 2026.

Qiyong Zhong, Mao Zheng, Mingyang Song, Xin Lin, Jie Sun, Houcheng Jiang, Xiang Wang, and Junfeng Fang. Sod: Step-wise on-policy distillation for small language model agents. *arXiv preprint arXiv:2605.07725*, 2026.

Appendix

A Notation . . . 14

B Step, Credit, and Cost of the Reweighting . . . 14

C Fisher Orthogonality and the First-Order Reward Change . . . 16

D Identification Constructions . . . 17

E Depth Trend of the Prefix Ratio . . . 18

F Batch Estimators and the Class Assignment . . . 18

G Reward of the Estimated Assignment . . . 19

H Additional Results . . . 20

I Implementation and Reproducibility . . . 21

J Case Study . . . 23

K Limitations . . . 23

A Notation

Symbol	Meaning	Symbol	Meaning
τ, x	rollout, task input	V_m, $\bar{V}_k$, $\hat{V}_k$	signal variance, class total, estimate
m, M	turn index, cap on turns	g_m, $\bar{g}_k$, $\hat{g}_k$	projected gain, class total, estimate
s_m, a_m, o_m	state, block, response at turn m	w_m, $w_m^\star$, $\hat{w}_k$	weight, calibrated weight, estimate
c_t, y_t	prefix and token at position t	c, ϵ, β	scale, floor, window rate
$\pi_\theta^{(w)}(\cdot \mid s_m)$	reweighted block distribution	ρ_m	prefix importance ratio
π_θ, π_T	student, teacher	ℓ_t, d_t, $\tilde{\ell}_m$, L	log-ratio, centered term, signal, bound
$R(\tau)$, $\eta(\pi)$, Q^{π_θ}, A^{π_θ}	reward, its expectation, value, advantage	$k(m)$, K, $\mathcal{K}$	class of a turn, class count, executed classes
μ, Δ, $\Delta_{\boldsymbol{w}}$	multiplier, trust-region radius, divergence of a profile	N, b_m, δ	rollouts in the window, baseline, failure probability

B Step, Credit, and Cost of the Reweighting

Centered signal and the chain rule. Fix a turn m and, throughout the proofs, write $p = \pi_\theta(\cdot \mid s_m)$ and $q = \pi_T(\cdot \mid s_m)$ for the two block distributions, $p^{(w)} = \pi_\theta^{(w)}(\cdot \mid s_m)$ for the reweighted distribution of Eq. (6), $Q_m(a) = \mathrm{E}[R \mid s_m, a_m = a]$ for the block value, and $\varepsilon(w) = 20w^2L^2$. At a position t with prefix c_t, the conditional mean of the token log-ratio is $\mathrm{E}_{y_t\sim\pi_\theta(\cdot|c_t)}[\ell_t] = \sum_y \pi_\theta(y \mid c_t) \log \pi_T(y \mid c_t)/\pi_\theta(y \mid c_t) = -\mathrm{KL}(\pi_\theta(\cdot \mid c_t) \,\|\, \pi_T(\cdot \mid c_t))$, so the term d_t of Eq. (1) has $\mathrm{E}[d_t \mid c_t] = 0$. The prefix $c_{t'}$ of a later position contains y_t, so for $t < t'$ the tower property gives $\mathrm{E}[d_t d_{t'} \mid s_m] = \mathrm{E}[d_t\, \mathrm{E}[d_{t'} \mid c_{t'}] \mid s_m] = 0$: the terms are martingale differences, $\mathrm{E}[\tilde{\ell}_m \mid s_m] = 0$, and $\mathrm{E}[\tilde{\ell}_m^2 \mid s_m] = \sum_t \mathrm{E}[d_t^2 \mid s_m]$, the sum of the per-position variances. The block distributions factorize over tokens, so $\log p(a) - \log q(a) = -\sum_t \ell_t$, and taking the expectation position by position gives $\mathrm{KL}(p \,\|\, q) = \mathrm{E}_{a\sim p} \sum_{t\in a} \mathrm{KL}(\pi_\theta(\cdot \mid c_t) \,\|\, \pi_T(\cdot \mid c_t))$, the chain rule of relative entropy, so the turn-m term of Eq. (2) is $w_m \mathrm{E}_\tau[\mathrm{KL}(p \,\|\, q)]$.

Policy-gradient form. Differentiating the position-t divergence $\mathrm{E}_{y\sim\pi_\theta(\cdot|c_t)}[\log \pi_\theta(y \mid c_t) - \log \pi_T(y \mid c_t)]$ with the score identity $\nabla_\theta \mathrm{E}_{p_\theta}[f] = \mathrm{E}_{p_\theta}[f\, \nabla_\theta \log p_\theta] + \mathrm{E}_{p_\theta}[\nabla_\theta f]$ gives $-\mathrm{E}[\ell_t \nabla \log \pi_\theta(y_t \mid c_t)] + \mathrm{E}[\nabla \log \pi_\theta(y_t \mid c_t)]$, and the second term vanishes since $\mathrm{E}_{y\sim\pi_\theta}[\nabla \log \pi_\theta(y \mid c_t)] = \nabla \sum_y \pi_\theta(y \mid c_t) = 0$. The divergence at the position is a function of c_t, so $\mathrm{E}[\mathrm{KL}(\pi_\theta(\cdot \mid c_t) \,\|\, \pi_T(\cdot \mid c_t))\, \nabla \log \pi_\theta(y_t \mid c_t) \mid c_t] = 0$ by the same identity, and ℓ_t can be replaced by d_t. Summing over the positions with the visitation of the prefixes kept fixed gives the first line of Eq. (3), and the second line is the policy-gradient theorem at the token level.

The reweighted distribution. For a distribution r on blocks and $w \geq 0$, maximizing $w\mathrm{E}_r[\tilde{\ell}_m] - \mathrm{KL}(r \,\|\, p)$ over r gives $r = p\, e^{w\tilde{\ell}_m}/Z_w$ with $Z_w = \mathrm{E}_{a\sim p}[e^{w\tilde{\ell}_m(a)}]$, since $w\mathrm{E}_r[\tilde{\ell}_m] - \mathrm{KL}(r \,\|\, p) = \log Z_w - \mathrm{KL}\Big(r \,\Big\|\, p\, e^{w\tilde{\ell}_m}/Z_w\Big)$ and relative entropy is nonnegative and vanishes at equality. The minimizer of the divergence to p at a fixed expected proxy advantage is therefore the reweighted distribution of Eq. (6), the closed form of a KL-constrained step whose reward is the proxy advantage.

Geometry of the step. Write $\Lambda(w) = \log \mathrm{E}_p[e^{w\tilde{\ell}_m}]$ for the log-partition function, so that $p^{(w)} = p\, e^{w\tilde{\ell}_m - \Lambda(w)}$ is the exponential family with natural parameter w and sufficient statistic $\tilde{\ell}_m$ (Amari, 2016). Standard exponential-family identities give $\Lambda'(w) = \mathrm{E}_{p^{(w)}}[\tilde{\ell}_m]$, which vanishes at $w = 0$, and $\Lambda''(w) = \mathrm{Var}_{p^{(w)}}(\tilde{\ell}_m) = V_m(w)$, the Fisher information of the family at w, with $V_m(0) = V_m$; $\Lambda'''(w)$ is the third central moment of $\tilde{\ell}_m$ under $p^{(w)}$. The divergence from the reweighted distribution back to p is

$$\mathrm{KL}\Big(p^{(w)} \,\Big\|\, p\Big) = \mathrm{E}_{p^{(w)}}\big[w\tilde{\ell}_m - \Lambda(w)\big] = w\Lambda'(w) - \Lambda(w),$$

$$\frac{d}{dw}\big[w\Lambda'(w) - \Lambda(w)\big] = w\Lambda''(w) = w\, V_m(w),$$

and the value at $w = 0$ is zero, so $\mathrm{KL}\big(p^{(w)} \,\big\|\, p\big) = \int_0^w s\, V_m(s)\, ds$, the integral of the Fisher information along the family.

Projection. The objective of Lemma 1, $\mathrm{E}_p[\sum_t (A_t - w d_t)^2]$ with $A_t = A^{\pi_\theta}(c_t, y_t)$, is a convex quadratic in w whose derivative $-2\mathrm{E}_p[\sum_t (A_t - w d_t) d_t]$ vanishes at $w^\star = \mathrm{E}_p[\sum_t d_t A_t]/\mathrm{E}_p[\sum_t d_t^2]$, and the denominator of $w^\star$ is V_m by the martingale paragraph. For the numerator, $A_t = Q^{\pi_\theta}(c_t, y_t) - \mathrm{E}[Q^{\pi_\theta}(c_t, y) \mid c_t]$ and the subtracted term is a function of c_t, so $\mathrm{E}[d_t A_t \mid c_t] = \mathrm{E}[d_t Q^{\pi_\theta}(c_t, y_t) \mid c_t]$ since $\mathrm{E}[d_t \mid c_t] = 0$; then $Q^{\pi_\theta}(c_t, y_t) = \mathrm{E}[R \mid c_t, y_t]$ and the tower property give $\mathrm{E}[d_t Q^{\pi_\theta}(c_t, y_t) \mid s_m] = \mathrm{E}[d_t R \mid s_m]$. Summing over the positions, the numerator is $\mathrm{E}[\tilde{\ell}_m R \mid s_m] = g_m$, which gives Eq. (4). The same computation with the block value gives $g_m = \mathrm{E}_p[\tilde{\ell}_m Q_m] = \mathrm{Cov}_p(\tilde{\ell}_m, Q_m)$, the form the paragraphs below use. For the Fisher reading, the tangent space of the block distributions at p is the space of centered functions with the inner product $\langle f, h\rangle = \mathrm{E}_p[fh]$. The tilt of Eq. (6) has score $\tilde{\ell}_m$ at $w = 0$, and the natural gradient of $r \mapsto \mathrm{E}_r[Q_m]$ at $r = p$ is $A_{\pi_\theta} = Q_m - \mathrm{E}_p Q_m$, so the inner product of the two directions is $\mathrm{E}_p[\tilde{\ell}_m A_{\pi_\theta}] = g_m$ and the squared length of the distillation direction is V_m.

Decomposition. With $w^\star$ the projection coefficient, the residual $r_t = A_t - w^\star d_t$ satisfies $\mathrm{E}_p[\sum_t d_t r_t] = \mathrm{E}_p[\sum_t d_t A_t] - w^\star \mathrm{E}_p[\sum_t d_t^2] = 0$ by the definition of $w^\star$, and substituting $A_t = w^\star d_t + r_t$ into the token-level policy gradient of Eq. (3) gives Eq. (5) by linearity, the first term being the distillation direction at weight $w^\star$ by the policy-gradient paragraph. The residual variance is $\mathrm{E}_p[\sum_t r_t^2] = \mathrm{E}_p[\sum_t A_t^2] - g_m^2/V_m$ by the same orthogonality, which is minimal over all coefficients at $w^\star$. The signal is therefore the control variate of the advantage with the least remaining variance, and the same identity with the return in place of the advantage gives the class-level fraction $1 - \kappa_k$ that sets the residual gate.

Credit. Write $x(a) = w\tilde{\ell}_m(a)$ for the exponent of the tilt, so that $p^{(w)}(a)/p(a) = e^{x(a)}/\mathrm{E}_p[e^x]$. At $w = 0$ the derivative of $e^x/\mathrm{E}_p e^x$ is $\tilde{\ell}_m(a)$, so for every bounded f, $\frac{d}{dw}\mathrm{E}_{p^{(w)}}[f]\big|_0 = \mathrm{E}_p[\tilde{\ell}_m f] = \mathrm{Cov}_p(\tilde{\ell}_m, f)$. With $f = A_{\pi_\theta}$ the covariance equals g_m, since a constant shift of f leaves a covariance unchanged. For the remainder, let $m_w = \mathrm{E}_p[e^x] \geq 1$ by Jensen's inequality since $\mathrm{E}_p[x] = 0$ and $|x| \leq wL$. Then

$$\frac{p^{(w)}}{p} - 1 - x = \frac{e^x - m_w - x\, m_w}{m_w} = \frac{(e^x - 1 - x) - (m_w - 1) - x(m_w - 1)}{m_w}.$$

The inequality $|e^x - 1 - x| \leq \frac{1}{2}x^2 e^{|x|}$ gives $|e^x - 1 - x| \leq \frac{1}{2}w^2L^2e^{wL}$, the same bound applies to $m_w - 1 = \mathrm{E}_p[e^x - 1 - x]$, and $|x(m_w - 1)| \leq wL \cdot \frac{1}{2}w^2L^2e^{wL}$; with $m_w \geq 1$ the three terms give $|p^{(w)}/p - 1 - x| \leq w^2L^2e^{wL}(1 + wL)$. For $wL \leq \frac{1}{2}$ the right side is at most $w^2L^2 \cdot e^{1/2} \cdot \frac{3}{2} \leq 20w^2L^2 = \varepsilon(w)$. Since $|A_{\pi_\theta}| \leq 1$, the reward remainder obeys $\big|\mathrm{E}_{p^{(w)}}[A_{\pi_\theta}] - wg_m\big| = \big|\mathrm{E}_p[(p^{(w)}/p - 1 - x)A_{\pi_\theta}]\big| \leq \varepsilon(w)$, which is the credit remainder used

below, and the same expansion gives $\mathrm{TV}(p^{(w)}, p) = \frac{1}{2}\mathrm{E}_p|p^{(w)}/p - 1| \le w\, V_m^{\mathrm{TV}} + \frac{1}{2}\varepsilon(w)$ with $V_m^{\mathrm{TV}} = \frac{1}{2}\mathrm{E}_p|\tilde{\ell}_m|$, the bound used below.

Cost. The geometry paragraph gives the remainder of the cost expansion in closed form as the integral

$$\mathrm{KL}\Big(p^{(w)} \,\Big\|\, p\Big) - \tfrac{1}{2}w^2 V_m = \int_0^w s\,\big(V_m(s) - V_m(0)\big)\,ds,$$

which the third cumulant controls. Under $p^{(s)}$ the signal lies in $[-L, L]$, so its central moments satisfy $|\tilde{\ell}_m - \mathrm{E}_{p^{(s)}}\tilde{\ell}_m| \le 2L$ and $|\Lambda'''(s)| \le 2L\,\mathrm{Var}_{p^{(s)}}(\tilde{\ell}_m) \le 2L^3$, the last step by Popoviciu's inequality $\mathrm{Var} \le (2L)^2/4$; hence $|V_m(s) - V_m(0)| \le 2L^3 s$ and the cost remainder is at most $\int_0^w 2L^3 s^2\,ds = \frac{2}{3}L^3 w^3$.

Bounds. Since the signal lies in $[-L, L]$ with zero mean, $V_m = \mathrm{E}_p[\tilde{\ell}_m^2] \le L^2$, and the gain is bounded through the variance as follows. Since $\mathrm{E}_p[\tilde{\ell}_m] = 0$, subtracting the constant $\frac{1}{2}$ from Q_m leaves the gain unchanged, $g_m = \mathrm{E}_p[\tilde{\ell}_m(Q_m - \frac{1}{2})]$, and $|Q_m - \frac{1}{2}| \le \frac{1}{2}$ gives $|g_m| \le \frac{1}{2}\mathrm{E}_p|\tilde{\ell}_m| = V_m^{\mathrm{TV}} \le \frac{1}{2}\sqrt{V_m}$ by Jensen's inequality. Hence $|w_m^\star|\sqrt{V_m} = |g_m|/\sqrt{V_m} \le \frac{1}{2}$: the calibrated weight is bounded by one half in Fisher units. Clipping d_t to $[-L_t, L_t]$ replaces it by $\bar{d}_t = \max\{-L_t, \min\{d_t, L_t\}\}$, whose conditional mean is $\mathrm{E}[\bar{d}_t - d_t \mid c_t]$ and is therefore bounded in absolute value by $\mathrm{E}[(|d_t| - L_t)_+ \mid c_t]$, the expected excess beyond the clip. The batch estimators are unbiased for the clipped signal, and their bias for the raw signal is at most that excess, summed over the block, times the bound on the centered return. For a class total, Cauchy–Schwarz over the turns of a rollout and Jensen's inequality give $|\bar{g}_k| \le \frac{1}{2}\mathrm{E}[\sum_{k(m)=k}|\tilde{\ell}_m|] \le \frac{1}{2}\sqrt{M\bar{V}_k}$, and $\bar{V}_k \le ML^2$.

C Fisher Orthogonality and the First-Order Reward Change

Throughout, the step of Proposition 1 has coordinates $\alpha\boldsymbol{w}$; we write w_m for αw_m in the proofs below and restore the factor α at the end, which turns the bound $LW\,D + 10L^2(2 + LW)\sum_m w_m^2$ into that of Eq. (7).

Chain rule for trajectories. The rollout distribution under $\pi_\theta^{(\boldsymbol{w})}$ is $P_{\boldsymbol{w}}(\tau) = \prod_m \pi_\theta^{(w_m)}(a_m \mid s_m)\,T(o_m \mid s_m, a_m)$ with the same environment kernel T at every $\boldsymbol{w}$. The kernels cancel in $\log P_{\boldsymbol{w}}/P_{\mathbf{0}} = \sum_m \log \pi_\theta^{(w_m)}(a_m \mid s_m)/\pi_\theta(a_m \mid s_m)$, and taking the expectation under $P_{\boldsymbol{w}}$ turn by turn through the tower property gives $\mathrm{KL}(P_{\boldsymbol{w}} \,\|\, P_{\mathbf{0}}) = \sum_m \mathrm{E}_{P_{\boldsymbol{w}}}[\mathrm{KL}\Big(\pi_\theta^{(w_m)}(\cdot \mid s_m) \,\Big\|\, \pi_\theta(\cdot \mid s_m)\Big)]$, the chain-rule form behind the divergence identity of Proposition 1.

Fisher orthogonality. The score of $P_{\boldsymbol{w}}$ with respect to w_m at $\boldsymbol{w} = \mathbf{0}$ is $\partial_{w_m}\log P_{\boldsymbol{w}}(\tau)|_{\mathbf{0}} = \tilde{\ell}_m(a_m)$, by the geometry paragraph of Appendix B with $\Lambda'(0) = 0$, and its conditional mean given s_m is zero. For $k < m$ the state s_m contains a_k, so the score of turn k is a function of s_m, and the tower property gives

$$F_{km} = \mathrm{E}_{\pi_\theta}\big[\tilde{\ell}_k(a_k)\,\tilde{\ell}_m(a_m)\big] = \mathrm{E}_{\pi_\theta}\big[\tilde{\ell}_k(a_k)\,\mathrm{E}[\tilde{\ell}_m(a_m) \mid s_m]\big] = 0,$$

while $F_{mm} = \mathrm{E}_{\pi_\theta}[\mathrm{E}[\tilde{\ell}_m^2 \mid s_m]] = \mathrm{E}_{\pi_\theta}[V_m(s_m)]$. The second-order expansion of the divergence in its parameters, $\mathrm{KL}(P_{\boldsymbol{w}} \,\|\, P_{\mathbf{0}}) = \frac{1}{2}\boldsymbol{w}^\top F \boldsymbol{w} + O(\|\boldsymbol{w}\|^3)$, then has vanishing cross terms, which is the divergence identity of Proposition 1. The same computation gives $\partial_{w_m}\eta(\pi_\theta^{(\boldsymbol{w})})|_{\mathbf{0}} = \mathrm{E}_{\pi_\theta}[g_m(s_m)]$ by the credit paragraph of Appendix B and the performance difference identity below. The natural gradient of the reward in the turn coordinates therefore has components $\mathrm{E}[g_m]/\mathrm{E}[V_m]$ (Amari, 1998), and with state-dependent weights the same computation applies pointwise and returns the per-state ratio of Eq. (4).

Performance difference identity. Let $V_{\pi_\theta}(s_m) = \mathrm{E}[R \mid s_m,\ \text{student continues}]$, with $V_{\pi_\theta}(s_{M+1}) = R(\tau)$ at the terminal state, and note $Q_m(a) = \mathrm{E}[V_{\pi_\theta}(s_{m+1}) \mid s_m, a_m = a]$ because the continuation after the response follows the student. For every rollout law π', the tower

property and a telescoping sum give

$$\mathrm{E}_{\tau\sim\pi'}\Big[\sum_m A_{\pi_\theta}(s_m,a_m)\Big] = \mathrm{E}_{\tau\sim\pi'}\Big[\sum_m \big(\mathrm{E}[V_{\pi_\theta}(s_{m+1}) \mid s_m,a_m] - V_{\pi_\theta}(s_m)\big)\Big]$$
$$= \mathrm{E}_{\tau\sim\pi'}\big[V_{\pi_\theta}(s_{M+1}) - V_{\pi_\theta}(s_1)\big] = \eta(\pi') - \eta(\pi_\theta),$$

the identity of Kakade & Langford (2002). With $\pi' = \pi_\theta^{(\boldsymbol{w})}$, conditioning on s_m and applying the credit paragraph of Appendix B, $\mathrm{E}[A_{\pi_\theta}(s_m,a_m) \mid s_m] = w_m g_m(s_m) + r_m(s_m)$ with $|r_m| \le \varepsilon(w_m)$. Writing $X(\tau) = \sum_m w_m g_m(s_m)$, which depends on the state sequence alone,

$$\eta(\pi_\theta^{(\boldsymbol{w})}) - \eta(\pi_\theta) = \mathrm{E}_{\tau\sim\pi_\theta^{(\boldsymbol{w})}}[X(\tau)] + \mathrm{E}_{\tau\sim\pi_\theta^{(\boldsymbol{w})}}\Big[\sum_m r_m(s_m)\Big],$$

and the second term is at most $\sum_m \varepsilon(w_m)$ in absolute value.

Coupling. Run a student rollout τ and a reweighted rollout τ' on one probability space with common environment seeds. At each turn the two blocks are drawn from the maximal coupling of $(p^{(w_m)}, p)$ at the current state, which produces different blocks with probability exactly $\mathrm{TV}(p^{(w_m)}, p)$. With that probability the coupling draws the two blocks from the residual distributions $(p^{(w)} - \min\{p^{(w)}, p\})/\mathrm{TV}$ and $(p - \min\{p^{(w)}, p\})/\mathrm{TV}$, and otherwise it draws one block from the overlap $\min\{p^{(w)}, p\}/(1-\mathrm{TV})$ for both. The rollouts coincide up to the first turn at which the coupled blocks differ, and every earlier turn is a turn of the student rollout, so the union bound and the total-variation bound of Appendix B give $\mathrm{P}(\tau' \ne \tau) \le \sum_m \mathrm{E}_{\tau\sim\pi_\theta}\big[\mathrm{TV}(p^{(w_m)}, p)\big] \le \sum_m \big(w_m \mathrm{E}_{\pi_\theta}[V_m^{\mathrm{TV}}(s_m)] + \frac{1}{2}\varepsilon(w_m)\big) \le D + 10L^2 \sum_m w_m^2$, with $D = \frac{1}{2}\sum_m w_m \mathrm{E}_{\pi_\theta}[\sqrt{V_m(s_m)}]$ by $V_m^{\mathrm{TV}} \le \frac{1}{2}\sqrt{V_m}$.

Bound. Write $W = \sum_m w_m$ for the total weight and $D = \frac{1}{2}\sum_m w_m \mathrm{E}_{\pi_\theta}[\sqrt{V_m(s_m)}]$, so that by the bounds of Appendix B, $|X| \le \sum_m w_m |g_m| \le \frac{1}{2}\sum_m w_m \sqrt{V_m} \le \frac{1}{2} LW$, and $X(\tau') = X(\tau)$ when the state sequences coincide. Hence $\big|\mathrm{E}_{\pi_\theta^{(\boldsymbol{w})}}[X] - \mathrm{E}_{\pi_\theta}[X]\big| \le \mathrm{E}\big[|X(\tau') - X(\tau)|\,\mathbf{1}\{\tau' \ne \tau\}\big] \le LW\,\mathrm{P}(\tau' \ne \tau) \le LW\,D + 10L^3 W \sum_m w_m^2$, and adding the remainder term of the identity bounds the left side of Eq. (7) by $LW\,D + 10L^3 W \sum_m w_m^2 + 20L^2 \sum_m w_m^2 = LW\,D + 10L^2(2 + LW)\sum_m w_m^2$, the explicit bound on the second-order term of Eq. (7), which we call rem; with the factor α restored it reads $\alpha^2[LW\,D + 10L^2(2+\alpha LW)\sum_m w_m^2]$. For the control by the divergence, Jensen's inequality gives $\mathrm{E}[\sqrt{V_m}] \le \sqrt{\mathrm{E}[V_m]}$ and Cauchy–Schwarz gives $\sum_m w_m \sqrt{\mathrm{E}[V_m]} \le \sqrt{M \sum_m w_m^2 \mathrm{E}[V_m]} = \sqrt{2M\Delta_{\boldsymbol{w}}}$ with $\Delta_{\boldsymbol{w}} = \frac{1}{2}\sum_m w_m^2 \mathrm{E}[V_m]$ the divergence of the profile, so $D \le \sqrt{M\Delta_{\boldsymbol{w}}/2}$. The whole right side is second order in the weights, since $D \le \frac{1}{2}LW$. □

D Identification Constructions

Response routing. Fix the task distribution, both policies, and the transition behavior up to s_m with $p \ne q$, so that $\tilde{\ell}_m$ is not constant under p, and let two environments differ only in the responses each block triggers at turn m and in the rewards downstream. Take the set $A = \{a : \tilde{\ell}_m(a) > 0\}$ of blocks the teacher prefers, which has $p(A) \in (0,1)$. In the first environment the response routes every block in A to a state whose continuation succeeds and every other block to one whose continuation fails, so $Q_m = 1$ on A and 0 elsewhere; in the second environment the routing swaps. Every statistic computed from teacher and student probabilities on the observed prefix takes the same value in the two environments. The gains differ, since in the first environment

$$g_m = \mathrm{E}_p\big[\tilde{\ell}_m Q_m\big] = p(A)\,\mathrm{E}_p[\tilde{\ell}_m \mid A] > 0,$$

and the swapped routing negates it. Mixing the two routings with weight λ gives $g_m(\lambda) = (1 - 2\lambda)g_m$, which moves through zero at $\lambda = \frac{1}{2}$, so weights built from policy statistics cannot determine the value or the sign of the gain; the magnitude bound $|g_m| \le \frac{1}{2}\sqrt{V_m}$ is the only information they carry.

Aggregation of the terminal return. Weights that are functions of $R(\tau)$ assign one value per trajectory. Under the mixed environment two rollouts can both end at $R = 1$ while passing at turn

m through states of opposite gain, one where the blocks the teacher prefers succeed and one where they fail. Every return-measurable weight then gives the two turns the same pull while their gains differ in sign. The product $\tilde{\ell}_m R$ is not a function of the return alone. At the first state it is positive on the successful rollouts, which carry the preferred blocks, and at the second state it is negative on them. Its average over the rollouts through each state therefore recovers the sign of the gain, which is the estimator of Eq. (8).

E Depth Trend of the Prefix Ratio

Many weighting rules accumulate the signal along the sampled prefix through the prefix importance ratio $\rho_m = \prod_{k \le m} \pi_T(a_k \mid s_k)/\pi_\theta(a_k \mid s_k)$ computed without clipping, which has $\mathrm{E}_{\pi_\theta}[\rho_m] = 1$ under student sampling while its expected logarithm falls with depth,

$$\mathrm{E}_{\pi_\theta}[\log \rho_m] = -\sum_{k \le m} \mathrm{E}_{\pi_\theta}\Big[\mathrm{KL}(\pi_\theta(\cdot \mid s_k) \,\|\, \pi_T(\cdot \mid s_k))\Big]. \tag{9}$$

The geometric mean of the ratio therefore falls with depth while its arithmetic mean remains one, so normalizing ρ_m within a rollout gives late turns a smaller share at the rate of policy separation, independent of their gain. By the chain rule of Appendix B the per-token divergences summed over a block equal the block divergence $\mathrm{KL}(p_k \,\|\, q_k)$ in expectation, so Eq. (9) reads $\mathrm{E}_{\pi_\theta}[\log \rho_m] = -\sum_{k \le m} \mathrm{E}_{\pi_\theta}[\mathrm{KL}(p_k \,\|\, q_k)]$. Taking logarithms in the definition of ρ_m turns the product into the sum

$$\log \rho_m = \sum_{k \le m} \sum_{t \in a_k} \log \frac{\pi_T(y_t \mid y_{<t}, x)}{\pi_\theta(y_t \mid y_{<t}, x)}.$$

Each summand carries the per-token divergence, since conditioning on the prefix gives

$$\begin{aligned} \mathrm{E}_{y_t \sim \pi_\theta}\left[\log \frac{\pi_T(y_t \mid y_{<t}, x)}{\pi_\theta(y_t \mid y_{<t}, x)} \,\middle|\, y_{<t}\right] &= \sum_y \pi_\theta(y \mid y_{<t}, x)\, \log \frac{\pi_T(y \mid y_{<t}, x)}{\pi_\theta(y \mid y_{<t}, x)} \\ &= -\,\mathrm{KL}(\pi_\theta(\cdot \mid y_{<t}, x) \,\|\, \pi_T(\cdot \mid y_{<t}, x)) \end{aligned}$$

by the definition of the divergence, and the tower property over prefixes and rollouts turns the sum of conditional expectations into Eq. (9).

The arithmetic mean stays at one, since conditioning on the prefix gives

$$\mathrm{E}_{y_t \sim \pi_\theta}\left[\frac{\pi_T(y_t \mid y_{<t}, x)}{\pi_\theta(y_t \mid y_{<t}, x)} \,\middle|\, y_{<t}\right] = \sum_y \pi_\theta(y \mid y_{<t}, x)\, \frac{\pi_T(y \mid y_{<t}, x)}{\pi_\theta(y \mid y_{<t}, x)} = \sum_y \pi_T(y \mid y_{<t}, x) = 1,$$

so the conditional expectation of ρ_m given its past equals its previous value, the product telescopes to $\mathrm{E}_{\pi_\theta}[\rho_m] = 1$ at every depth, and ρ_m is a positive martingale. Jensen's inequality applied to the strictly concave logarithm gives $\mathrm{E}[\log(\pi_T/\pi_\theta)] < \log \mathrm{E}[\pi_T/\pi_\theta] = 0$ strictly wherever the token distributions differ, so every summand of Eq. (9) is nonpositive and strictly negative at the turns the teacher would alter, while turns with coincident distributions contribute zero.

F Batch Estimators and the Class Assignment

Unbiasedness. For a turn m of a rollout, $\mathrm{E}[\tilde{\ell}_m R \mid s_m] = g_m(s_m)$ by the projection paragraph of Appendix B, and $\mathrm{E}[\tilde{\ell}_m^2 \mid s_m] = V_m(s_m)$ by definition. Since a baseline b_m is a function of s_m, $\mathrm{E}[\tilde{\ell}_m b_m \mid s_m] = b_m \mathrm{E}[\tilde{\ell}_m \mid s_m] = 0$, so $\mathrm{E}[\tilde{\ell}_m (R - b_m) \mid s_m] = g_m(s_m)$ as well. The class $k(m)$ is a function of s_m, so summing over the turns of a rollout with $k(m) = k$ and taking the expectation over rollouts gives $\mathrm{E}[\sum_{k(m)=k} \tilde{\ell}_m (R - b_m)] = \mathrm{E}[\sum_{k(m)=k} g_m(s_m)] = \bar{g}_k$ and $\mathrm{E}[\sum_{k(m)=k} \tilde{\ell}_m^2] = \bar{V}_k$, so the averages of Eq. (8) over independent rollouts are unbiased for the class totals. The ratio $\hat{w}_k$ equals $\sum \tilde{\ell}_m (R_i - b_m) / \sum \tilde{\ell}_m^2$ over the turns of the class, the least-squares slope through the origin of the centered return on the signal.

Residual. For a token y_t of turn m, $\mathrm{E}[(R - b_m) \nabla \log \pi_\theta(y_t \mid c_t)] = \mathrm{E}[A_t \nabla \log \pi_\theta(y_t \mid c_t)]$ by the policy-gradient theorem with a state baseline. For the subtracted part, $\mathrm{E}[\tilde{\ell}_m \nabla \log \pi_\theta(y_t \mid c_t)] = \mathrm{E}[d_t \nabla \log \pi_\theta(y_t \mid c_t)]$, since the terms $d_{t'}$ with $t' < t$ are functions of c_t and vanish against

the score, and the terms with $t' > t$ vanish by $\mathrm{E}[d_{t'} \mid c_{t'}] = 0$ with $c_{t'}$ containing y_t. Hence $\mathrm{E}[\hat{r}_m \nabla \log \pi_\theta(y_t \mid c_t)] = \mathrm{E}[(A_t - \hat{w}_{k(m)} d_t) \nabla \log \pi_\theta(y_t \mid c_t)]$, the residual policy gradient of Eq. (5) at the estimated weight.

State-measurable classes. If the class of a turn depended on its own block, through the exit status of the current call, the weight $w_{k(m)}$ would be a function of (s_m, a_m). The proxy advantage of the turn would then become $w_{k(m)} \tilde{\ell}_m$ with $\mathrm{E}[w_{k(m)} \tilde{\ell}_m \mid s_m] \neq 0$ in general, and the single-sample estimates would pick up the covariance of the weight with the signal instead of the gain. Classes defined by the previous response, which is part of s_m, keep the weight a function of the state and the estimates unbiased.

Concentration. Write $G_i = \sum_{m \in \tau_i,\, k(m)=k} \tilde{\ell}_m (R_i - b_m)$ and $S_i = \sum_{m \in \tau_i,\, k(m)=k} \tilde{\ell}_m^2$ for the contributions of rollout i, which are independent across rollouts and lie in $[-ML, ML]$ and $[0, ML^2]$ since a rollout has at most M turns, $|\tilde{\ell}_m| \leq L$, and $|R - b_m| \leq 1$. Hoeffding's inequality for a mean of N independent variables with range r gives $\mathrm{P}(|\bar{Z} - \mathrm{E}\bar{Z}| \geq t) \leq 2\exp(-2Nt^2/r^2)$; with $r = 2ML$ and $t = MLt_N$, and with $r = ML^2$ and $t = ML^2 t_N$, where $t_N = \sqrt{2\log(4K/\delta)/N}$, each bound is at most $\delta/(2K)$. The estimation window weights the batches geometrically, and the same bounds apply with N replaced by the effective sample size of the weights (Kong, 1992; Elvira et al., 2022).

The class assignment. Write $v_k = \alpha w_k \geq 0$ for the step of class k and restrict the profile of Proposition 2 to weights that are constant on each class. The first-order reward of Eq. (7) is then $\mathrm{E}[\sum_m v_{k(m)} g_m(s_m)] = \sum_k v_k \bar{g}_k$ and the divergence is $\frac{1}{2}\sum_k v_k^2 \bar{V}_k$, so the program is

$$\max_{\boldsymbol{v} \geq 0} \sum_k v_k \bar{g}_k \qquad \text{subject to} \qquad \tfrac{1}{2} \sum_k v_k^2 \bar{V}_k \leq \Delta,$$

a linear function over a convex set, the orthant intersected with an ellipsoid, so the Karush–Kuhn–Tucker conditions are necessary and sufficient (Bonnans et al., 2006). With multipliers $\mu \geq 0$ for the radius and $\lambda_k \geq 0$ for the orthant, stationarity reads $\bar{g}_k - \mu v_k \bar{V}_k + \lambda_k = 0$. On a class with $\bar{g}_k \leq 0$ every positive step spends divergence and lowers the objective, so $v_k = 0$. On a class with $\bar{g}_k > 0$ the step is positive, $\lambda_k = 0$, and $v_k = \bar{g}_k/(\mu \bar{V}_k)$; the constraint binds since the objective increases along $\boldsymbol{v}^\star$, and complementary slackness fixes μ through $\frac{1}{2}\mu^{-2} \sum_{\bar{g}_k > 0} \bar{g}_k^2/\bar{V}_k = \Delta$. Hence $w_k = \bar{g}_k/(\alpha\mu\bar{V}_k)$ on the classes with positive gain, the ratio of the class totals up to the global scale $1/(\alpha\mu)$, which is the class-level statement of the main text. The same argument with one class per state returns Proposition 2. □

G Reward of the Estimated Assignment

Error of the executed weight. For a class in $\mathcal{K}$ the executed weight satisfies

$$\frac{\hat{g}_k}{\hat{V}_k} - \frac{\bar{g}_k}{\bar{V}_k} = \frac{\hat{g}_k - \bar{g}_k}{\hat{V}_k} + \bar{g}_k \Big(\frac{1}{\hat{V}_k} - \frac{1}{\bar{V}_k} \Big) = \frac{\hat{g}_k - \bar{g}_k}{\hat{V}_k} + \frac{\bar{g}_k \, (\bar{V}_k - \hat{V}_k)}{\hat{V}_k \bar{V}_k},$$

so $|\hat{w}_k - \bar{g}_k/\bar{V}_k| \leq |\hat{g}_k - \bar{g}_k|/\hat{V}_k + |\bar{g}_k| |\hat{V}_k - \bar{V}_k|/(\hat{V}_k \bar{V}_k)$. The miscalibration of the class in nats is $\frac{1}{2} c^2 (\hat{w}_k - \bar{g}_k/\bar{V}_k)^2 \bar{V}_k$, and $(a+b)^2 \leq 2a^2 + 2b^2$ bounds it by the sum of a gain-side term $c^2 (\hat{g}_k - \bar{g}_k)^2 \bar{V}_k / \hat{V}_k^2$ and a variance-side term $c^2 \bar{g}_k^2 (\hat{V}_k - \bar{V}_k)^2 / (\hat{V}_k^2 \bar{V}_k)$. The ablations report the two terms with the whole-run estimates for the unknowns; they bound the miscalibration and do not partition it, and offsetting errors, a gain overestimate against an oversized variance estimate, stay invisible to the split.

Excluded classes. On the event of Proposition 3 a class with $\hat{V}_k < \epsilon$ has $\bar{V}_k \leq \hat{V}_k + \epsilon/2 \leq \frac{3}{2}\epsilon$, and at a weight clipped to $w_m L \leq \frac{1}{2}$ it would gain at most $\frac{1}{2L}|\bar{g}_k| \leq \frac{1}{4L}\sqrt{M\bar{V}_k} \leq \sqrt{3M\epsilon/2}/(4L)$ in first-order reward by the bounds paragraph of Appendix B, which is the loss per excluded class stated in Section 4.2.

Proposition 3 (First-order reward of the estimated assignment)**.** *Let $\mathcal{K}$ be the classes with $\hat{V}_k \geq \epsilon$ and $\hat{g}_k > 0$, estimated from N independent rollouts among K classes, let $w_m = c\,\hat{w}_{k(m)}$ on those*

classes with $w_m L \le \frac{1}{2}$ and 0 elsewhere, and let $N \ge 8M^2L^4\epsilon^{-2}\log(4K/\delta)$. With probability at least $1-\delta$ over the rollouts, the first-order term of Proposition 1 obeys

$$\sum_{k\in\mathcal{K}} c\,\hat{w}_k\,\bar{g}_k \;\ge\; c\sum_{k\in\mathcal{K}}\frac{\bar{g}_k^2}{\bar{V}_k} \;-\; c\,|\mathcal{K}|\,\frac{M^2L^2}{2}\Big(\frac{1}{\epsilon}+\frac{ML^2}{\epsilon^2}\Big)\sqrt{\frac{2\log(4K/\delta)}{N}}. \tag{10}$$

Proof of Proposition 3. Write $t_N = \sqrt{2\log(4K/\delta)/N}$ for the deviation scale of the window, so that the condition $N \ge 8M^2L^4\epsilon^{-2}\log(4K/\delta)$ is $ML^2t_N \le \epsilon/2$. The first-order term of Proposition 1 under class weights is $\sum_k c\hat{w}_k\bar{g}_k$ by the class paragraph of Appendix F, and the transfer to η adds the second-order remainder of Appendix C. Classes outside $\mathcal{K}$ have weight 0, so the first-order term is $c\sum_{k\in\mathcal{K}}(\hat{g}_k/\hat{V}_k)\bar{g}_k = c\sum_{k\in\mathcal{K}}\bar{g}_k^2/\bar{V}_k - c\sum_{k\in\mathcal{K}}(\bar{g}_k/\bar{V}_k - \hat{g}_k/\hat{V}_k)\bar{g}_k$. By the concentration paragraph of Appendix F, each of the $2K$ events $\{|\hat{g}_k - \bar{g}_k| < MLt_N\}$ and $\{|\hat{V}_k - \bar{V}_k| < ML^2t_N\}$ fails with probability at most $\delta/(2K)$, so all of them occur with probability at least $1-\delta$. On that event, $\bar{V}_k \ge \hat{V}_k - ML^2t_N \ge \epsilon/2$ for $k \in \mathcal{K}$ by the floor and the assumption $ML^2t_N \le \epsilon/2$, and the error paragraph with $|\bar{g}_k| \le \frac{1}{2}\sqrt{M\bar{V}_k} \le \frac{1}{2}ML$ gives $|\hat{w}_k - \bar{g}_k/\bar{V}_k| \le MLt_N/\epsilon + \frac{1}{2}ML\cdot ML^2t_N/(\epsilon\cdot\epsilon/2) = MLt_N(1/\epsilon + ML^2/\epsilon^2)$. Multiplying by $c|\bar{g}_k| \le \frac{1}{2}cML$ and summing over $\mathcal{K}$ gives the second term of Eq. (10). □

H Additional Results

Tables 4 and 5 repeat the comparison with a Llama-3.1-70B-Instruct teacher and a Llama-3.1-8B-Instruct student (Grattafiori et al., 2024), so the calibration argument carries across model families. Figure 6 sweeps each calibration knob of Table 6 across its working range on ALFWorld.



Figure 5: Explained share κ_k of the return by the centered signal, per environment and state type, averaged over the run with the Qwen3-8B student. Cells without a class are blank, since the shorter environments end before the latest call-index groups. The share is lowest on the first call of a rollout and highest on the calls that follow a failed call, and the residual step is on in the cells below one half.

Table 4: Task performance on ALFWorld, Search-QA, and WebShop for a Llama-3.1-70B-Instruct teacher distilled into a Llama-3.1-8B-Instruct student, under the protocol, columns, and marks of Table 1.

Method	ALFWorld							Search-QA								WebShop	
	Pick	Look	Clean	Heat	Cool	Pick2	Avg	NQ	Triv.	Pop	Hotp.	2Wk	MuS.	Bam.	Avg	Score	SR
Llama-3.1-70B (Teacher)	98.3	92.2	99.4	95.7	95.2	92.9	96.1	66.4	80.3	62.8	66.8	69.7	51.5	67.2	66.4	93.4	85.3
Llama-3.1-8B-Instruct (Student)																	
Zero-Shot	45.0	38.9	34.8	29.6	29.5	25.9	34.3	29.4	52.4	29.9	24.6	28.6	2.5	31.4	28.4	33.5	12.2
GRPO (Shao et al., 2024)	88.3	57.8	85.8	69.6	66.7	56.5	73.0	39.1	63.0	39.5	36.5	41.3	11.8	36.3	38.2	77.1	61.6
Vanilla OPD (Agarwal et al., 2024)	90.8	63.3	90.3	75.7	73.3	63.5	78.2	41.5	64.6	41.6	39.3	43.8	14.4	39.0	40.6	80.2	65.7
SOD (Zhong et al., 2026)	91.7	65.6	92.3	77.4	75.2	64.7	79.9	42.9	65.9	43.1	42.2	46.6	17.5	42.2	42.9	81.7	67.7
IW-OPD (Xie et al., 2026)	91.7	62.2	89.7	75.7	72.4	62.4	77.8	42.6	65.4	43.1	38.8	43.1	13.6	38.2	40.7	79.9	65.1
TCOD (Wang et al., 2026a)	91.7	65.6	91.6	78.3	76.2	65.9	80.1	41.9	64.5	41.9	39.1	44.0	14.3	39.4	40.7	80.5	66.1
RLSD (Yang et al., 2026a)	91.7	65.6	92.3	77.4	76.2	65.9	80.1	45.9	**68.3**	**46.1**	44.1	48.4	19.4	44.0	45.2	84.1	68.1
TACT	**95.0**	**73.3**	**94.8**	**85.2**	**83.8**	**71.8**	**85.7**	**46.0**	**68.3**	46.0	**45.6**	**50.0**	**21.0**	**47.4**	**46.3**	**85.7**	**73.2**

Table 5: ScienceWorld and τ-bench for the same Llama pair, under the protocol, columns, and marks of Table 2.

Method	ScienceWorld		τ-bench	
	Score	SR	pass^1	pass^4
Llama-3.1-70B (Teacher)	88.5	85.2	55.4	40.6
Llama-3.1-8B-Instruct (Student)				
Zero-Shot	33.9	9.6	9.6	5.3
GRPO (Shao et al., 2024)	56.8	39.6	19.5	12.9
Vanilla OPD (Agarwal et al., 2024)	60.1	43.9	24.1	16.6
SOD (Zhong et al., 2026)	61.5	45.7	26.8	18.8
IW-OPD (Xie et al., 2026)	59.6	43.2	24.5	16.9
TCOD (Wang et al., 2026a)	61.6	45.8	25.4	17.6
RLSD (Yang et al., 2026a)	61.3	45.4	27.4	19.5
TACT	**67.6**	**53.4**	**33.5**	**22.1**

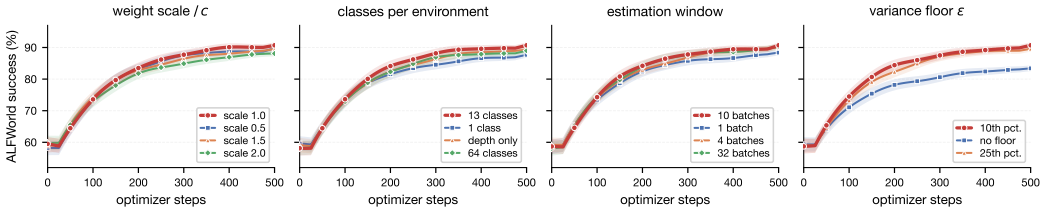

Figure 6: Hyperparameter sensitivity of TACT measured on ALFWorld training runs with the Qwen3-8B student. Each panel sweeps one calibration knob with every other setting at its default value, with rolling means over 5 seeds and one-standard-deviation bands. The weight-scale panel reports its values as multiples of the scale c. Success is least sensitive near the unit multiple and falls when the classes collapse to one per environment or split beyond the sample size of a window. It gains little beyond the effective window of ten batches that the rate $\beta = 0.9$ gives and loses the most without the variance floor.

I IMPLEMENTATION AND REPRODUCIBILITY

Table 6 records the training and calibration settings shared across environments, and Table 7 records what each environment supplies. The class of a turn is the triple of the index of its tool call, the exit status of the previous call, and the verifier output on the previous block. The index is grouped as the first call, the second to fourth, the fifth to twelfth, and later calls, and the exit status is success or failure. The verifier output is accept or reject where ALFWorld, ScienceWorld, and τ-bench provide a verifier and takes a single value on WebShop and Search-QA. The first turn of a rollout lacks a previous call and forms a class of its own, so the three verified environments have at most

thirteen classes and the other two at most seven. The baseline b_m is the mean return of the batch on the environment. The ablation variants of Table 3 replace one ingredient each. The variant w/o centering uses the raw block log-ratio $\sum_t \ell_t$ in place of $\tilde{\ell}_m$ in both sums of Eq. (8). The variant w/ policy gap sets $\hat{g}_k$ to the mean block divergence of the class, and w/ rollout sign replaces $\tilde{\ell}_m(R_i - b_m)$ by $R_i - b_m$, the sign rule of RLSD pooled by class. The variant w/o classes uses one class per environment, and w/ accum. ratio replaces $\hat{V}_k$ by the class mean of the normalized prefix ratio ρ_m. The variant w/o floor removes ϵ, and w/o residual step sets $\hat{r}_m = 0$ on every turn. The variant w/ fixed OPD + RL sum uses a constant OPD weight c and adds the reinforcement term with advantage $R - b_m$ on every turn, the structure of SOD.

Table 6: Training and calibration settings shared across all five environments.

Setting	Value	Setting	Value
teacher / students	Qwen3-32B / 4B / 8B	classes per environment K	$\leq$13 ($\leq$7 without a verifier)
optimizer	AdamW	window rate β	0.9
learning rate	1×10^{-6}	floor ϵ	10th percentile of $\hat{V}_k$
rollouts per batch	128	weight scale c	0.125 ($= 1/(2L)$)
sampling temperature	1.0	residual gate $\kappa_{\min}$	0.5, advantage $(R - b_m) - \hat{w}_k \tilde{\ell}_m$
optimizer steps	500	baseline b_m	batch mean return per environment
log-ratio clip L	4 (block level)	executed-weight clip	$w_m L \leq \frac{1}{2}$

We record backbone checkpoints, teacher configuration and hint format, the class definition and call-index groups per environment, optimizer settings, rollout budgets per environment, the measured distribution from which the floor percentile is taken, and the seeds. Code and configurations are available at `https://anonymous.4open.science/r/TACT_ICLR-BD0D/README.md`.

Table 7: Environment records: horizon caps, tool interfaces, the privileged hint the teacher reads, the turn-level feedback each environment logs for the classes, and evaluation sizes.

Environment	Max turns	Tool	Teacher hint	Logged feedback	Evaluation
ALFWorld	30	text actions	expert plan	status, verifier	134 unseen games
ScienceWorld	30	text actions	gold actions	status, verifier	270 test variations
WebShop	15	site actions	target attributes	status	500 test instructions
Search-QA	4	retrieval queries	gold passages	status	7 test splits
τ-bench	30	domain APIs	reference actions	status, verifier	115 retail, 50 airline

Table 8: Computation per optimizer step relative to vanilla OPD, counted from the update rule of each method under the shared-rollout protocol. Student generations count the rollouts and every continuation a method draws, teacher passes count the forward passes that score the sampled positions, and the last two columns list the loss terms and the state a method adds.

Method	Student generations	Teacher passes	Added loss term	Extra state
Vanilla OPD	1.00	1.00	none	none
SOD	1.00	1.00	GRPO term	one scalar per step
IW-OPD	1.00	1.00	none	one prefix sum per rollout
TCOD	1.00	1.00	none	fixed turn schedule
RLSD	1.00	1.00	none	one sign per rollout
GRPO	1.00	0	none	group statistics
TACT	1.00	1.00	gated residual term	three scalars per class

The matching loss evaluates the full-vocabulary KL at every position from the cached logits, so the update covers teacher modes the student rarely samples, and the sum over turns runs over the visited turns of each episode. The added computation of TACT comprises the per-position divergence, which the loss already evaluates, the centered log-ratio at the sampled token, and two scalar sums per class, without generation beyond the rollouts of on-policy distillation.

J CASE STUDY

The case is a Search-QA turn whose state type is the second call after an empty retrieval, taken from a training batch of the Qwen3-8B student at step 300. The question asks which river flows through the birthplace of a named composer, and two rollouts reach states of this type on it. The student's habitual query repeats the question verbatim, and the teacher assigns it a lower probability than the student does, so its centered signal is $\tilde{\ell}_m = -0.62$; the retrieval returns nothing again and the rollout fails, $R = 0$. The other rollout's query names the birthplace found in the first response and asks for its river, which the teacher prefers, $\tilde{\ell}_m = +0.81$; the retrieval returns the passage and the rollout succeeds, $R = 1$. With the batch mean return $b = 0.44$, the two turns contribute $(-0.62)(-0.44) = 0.27$ and $(0.81)(0.56) = 0.45$ to the gain of the class and 0.38 and 0.66 to its variance, both positive, and the class weight over the window is $\hat{w}_k = 0.71$. The explained share of the class is 0.52, above the gate, so the residual step is off for these turns and the update pulls the student toward the teacher's rewriting at the calibrated weight. On the first call of the same rollouts, where the explained share is 0.09, the weight is 0.08 and the residual step carries the return.

K LIMITATIONS

The analysis of Section 3 is first order in the reward and second order in the divergence, so it describes one update whose step stays small relative to the number of turns, and the effect of moved states enters only through the bounded second-order remainder. The estimates of Section 4 require a bounded return and a teacher that places positive probability wherever the student does, and their unbiasedness rests on the centering of the signal, which clipping perturbs by at most the expected excess beyond the clip (Appendix B). The classes must be functions of the state, which excludes the exit status of the current call from the class definition and leaves the resolution of the calibration at the level of state types instead of individual states. The residual step reintroduces the variance of the return on the classes where it is on, reduced by the explained share but not removed, and the gate $\kappa_{\min}$ stays at one half in every experiment. The guarantees of Proposition 3 carry Hoeffding constants that are conservative at the horizons of the experiments, so the experiments check the estimates by the measured miscalibration instead of the bound.