DIVERGENCE-CALIBRATED TURN IMPORTANCE FOR MULTI-TURN AGENTIC ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) has become a prevailing recipe for transferring multi-turn agentic capabilities from large teachers to small deployable students. However, effective multi-turn distillation remains challenging: OPD supervises at the token level, whereas an agentic decision is made by a complete turn, so token-level supervision can flatten turn-level decisions. Existing methods have recognized the multi-turn agentic property, but they either introduce environment rewards that pure OPD lacks or aim to skip unnecessary supervision, rebalance the loss budget, and keep supervision reliable. **This leaves an open question: how should supervision be allocated across turns by their importance when no environment reward is available?** We propose Divergence-Calibrated Turn Importance for Multi-Turn Agentic On-Policy Distillation (DT-OPD) to answer this question. We demonstrate that, under natural-gradient updates, the teacher-student divergence drift induced by a sampled response is determined by the response's stop-gradient negated divergence score and by its relative divergence. Our analysis lifts the divergence from the token level to the turn level, aligning it with the unit at which the agent acts. DT-OPD therefore turns the drift into a practical turn-level teacher-student divergence proxy and redistributes supervision across turns accordingly. Extensive experiments on ALFWorld, Search-QA, and WebShop, distilling Qwen3 and Qwen2.5 students from 0.6B to 4B using GRPO teachers up to Qwen3-30B-A3B, show that DT-OPD achieves the best overall performance across the three benchmarks and all student scales, and ablation studies and training dynamics analysis further confirm that both of our core contributions are essential to this performance.

1 INTRODUCTION

Large language model (LLM) agents have advanced from one-off question answering to multi-turn interactive decision making—observing, acting, and invoking tools across many turns (Yao et al., 2022b; Dong et al., 2026; Deng et al., 2023; Xi et al., 2025)—and increasingly evolve themselves through memory, skills, and recursive self-improvement (Wang et al., 2023; Yu et al., 2026; Chen et al., 2026b). However, these capabilities are realized by very large models, whose inference cost and deployment overhead motivate transferring them to smaller students (Mansourian et al., 2025). On-policy distillation (OPD) has become a prevailing recipe for this transfer: the student generates its own trajectories, and the teacher supervises its output distribution on those self-generated states, with no extrinsic reward (Gu et al., 2024; Agarwal et al., 2024; Lu & Lab, 2025; Li et al., 2026b).

Inspired by this self-generated-state supervision, a series of work tailored to multi-turn agentic tasks have been proposed in recent years (Wang et al., 2026; Liao et al., 2026; Sun et al., 2026; Lu et al., 2026). However, OPD faces a fundamental challenge in this setting: it supervises every token through token-level KL matching (Agarwal et al., 2024; Li et al., 2026b), whereas the atomic unit of an agentic decision is the complete turn (Yao et al., 2022b), so token-level supervision flattens turn-level decisions.

Motivated by this mismatch, recent work has moved supervision toward turns of uneven importance (Zhong et al., 2026; Tan et al., 2026). One strand weights turns, but draws the weights from environment rewards: ATOD places turn-level gates in a reward-annealed objective (Tan et al., 2026), SOD trains its turn weights alongside a GRPO term (Zhong et al., 2026), and reward-side structures

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.

sharpen turn-level supervision (Wei et al., 2025; Wen et al., 2024); without environment rewards, that signal is absent. A complementary strand reshapes supervision within OPD: it decides per turn whether the teacher is needed (Zhou et al., 2026a), or redistributes the loss across depth and length (Wang et al., 2026; Zhou et al., 2026b); the aims are to skip unnecessary supervision, rebalance the loss budget, and keep supervision reliable, not to judge how important a turn is. **This leaves an open question: how should each complete turn be weighted by its importance in this distillation, where no environment reward is available?**

We find that the teacher-student divergence at a turn measures the importance of that complete turn. Specifically, we notice that: (i) the teacher-student divergence already provides an intrinsic signal: large where the student still deviates from the frozen teacher, small where the two already agree; (ii) each completed turn is the effective unit that changes the agent's next decision. We therefore adopt it as the measure of a turn's importance, and let the OPD update decide how that importance becomes a weight. We demonstrate that the teacher-student divergence drift induced by a sampled response is determined by the stop-gradient negated divergence score of the response and by the response's relative divergence. This motivates Divergence-Calibrated Turn Importance for Multi-Turn Agentic On-Policy Distillation (DT-OPD), an algorithm that needs no environment reward and weights each complete turn by a group-relative proxy for the relative divergence. With strictly positive weights that are monotone in the score and average to one within each prompt, DT-OPD moves the distillation budget toward the turns the update can still move, without adding supervision or changing how much is distilled.

The following contributions are made in the course of this work:

- We provide a turn-level theoretical analysis of teacher-student divergence drift in multi-turn agentic OPD. By showing that the drift is determined by the stop-gradient negated divergence score and by the relative divergence, our analysis establishes that the relative divergence calibrates the importance of each turn.
- We propose DT-OPD, a lightweight method that weights each complete turn by a group-relative score built from its length-normalized teacher-student divergence. Because those weights are recomputed from the current distributions at every step, DT-OPD follows the divergence as the student changes instead of fixing a schedule in advance.
- We evaluate on ALFWorld, Search-QA, and WebShop, distilling Qwen3 and Qwen2.5 students from 0.6B to 4B with GRPO teachers up to Qwen3-30B-A3B. DT-OPD improves the per-benchmark average over the strongest baseline at every student scale, by up to 3.1 points on ALFWorld, 0.8 on Search-QA, and 4.4 on WebShop. Further analysis, including ablation studies and training dynamics analysis, confirms that both of our contributions are vital to this performance.

2 Related Work

Multi-turn Agents LLM agents that interleave reasoning with action and call external tools on their own have become a core paradigm for interactive decision making (Yao et al., 2022b; Schick et al., 2023). The paradigm now spans retrieval-augmented question answering (Jin et al., 2025), web interaction (Yao et al., 2022a), embodied textual environments (Shridhar et al., 2020), and repository-level software engineering (Yang et al., 2024), with surveys organizing the field through methodology-centered taxonomies, multimodal embodiment, and multi-turn task settings (Luo et al., 2025; Durante et al., 2024; Dong et al., 2026). Frontier models push this capability further: Kimi K3 and GLM-5 extend frontier-scale agentic execution to multi-turn settings (Team et al., 2026; Zeng et al., 2026). However, these capabilities are costly at frontier scale, so transferring them from large teachers into small, deployable students without reward annotation becomes the central problem of agent deployment (Zhong et al., 2026).

From Token-Level to Turn-Level On-Policy Distillation OPD trains a student on its own trajectories with dense token-level KL matching (Agarwal et al., 2024; Li et al., 2026b; Song & Zheng, 2026). In agentic tasks the environment reacts only after a complete action, so the turn, not the token, is the atomic decision unit. One class weights turns with environment rewards: ATOD gates turns by disagreement and entropy under reward annealing, SOD weights turns by divergence ratios

with a GRPO term, and reward-side work sharpens credit (Tan et al., 2026; Zhong et al., 2026; Wei et al., 2025; Wen et al., 2024). A second class moves where supervision lands: Step-OPD weights turns equally, TCOD expands the distilled span by curriculum, and TurnOPD blends length share with uniform weight (Sun et al., 2026; Wang et al., 2026; Zhou et al., 2026b). A third class asks the teacher when to teach: Guided-OPD mixes teacher turns under annealing, FutureBridge-OPD validates guidance at high-disagreement states, SAGE-OPD labels turns Skip, Weak, or Strong, and DASH-OPD switches executors with hysteresis (Li et al., 2026a; Chen et al., 2026a; Zhou et al., 2026a; Xia et al., 2026). A fourth class manages supervision cost: ReOPD replays reliability-weighted prefixes, and f-OPD downweights stale samples (Liao et al., 2026; Chen et al., 2026c). The last three settle that weight outside the rollout and the first needs rewards OPD lacks, so how much a complete turn should weigh stays open; we weight turns by their own divergence.

3 PRELIMINARY

Before formally elaborating the DT-OPD method, this section first introduces relevant preparatory knowledge to lay the foundation for the subsequent theoretical derivation and model construction.

3.1 MULTI-TURN AGENTIC INTERACTION SETUP

We consider a multi-turn agentic setting, where an agent policy $\pi_\theta(\cdot \mid s)$ interacts with an environment over T turns. At each turn $t \in \{0, \dots, T-1\}$, the agent observes a state $s_t \in \mathcal{S}$, e.g., language messages, tool outputs, or webpage snapshots, and produces a textual response $a_t = (y_1, \dots, y_{|a_t|}) \in \mathcal{A}_t \subset V^{\leq n}$, e.g., free-form text, a tool call with arguments, or an interface selection, where V is the LLM vocabulary and n is the maximum output length. Given a prompt s_0, an episode yields a trajectory $\tau = \{(s_0, a_0), \dots, (s_{T-1}, a_{T-1})\}$, which is sampled from $P_\theta(\cdot \mid s_0) = \prod_{t=0}^{T-1} \pi_\theta(\cdot \mid s_t)$ under the Markov decision process assumption, conditioned on the initial state s_0 drawn from a distribution $\mathcal{D}$. The environment reacts only after a complete response is generated, making the turn, rather than an individual token, the effective interaction unit.

3.2 AGENTIC ON-POLICY DISTILLATION OBJECTIVE

OPD generalizes from a single response to a whole episode: the multi-turn objective $\mathcal{L}_{\mathrm{OPD}}(\theta)$ is the trajectory-level reverse KL between the student's and the teacher's trajectory distributions (Zhou et al., 2026b; Li et al., 2026b), with the teacher queried only at student-visited states:

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{s_0 \sim \mathcal{D}}\Big[D_{\mathrm{KL}}\big(P_\theta(\cdot \mid s_0) \,\|\, P_T(\cdot \mid s_0)\big)\Big], \tag{1}$$

where $s_0 \sim \mathcal{D}$ is the initial state, and $P_\theta(\cdot \mid s_0)$ and $P_T(\cdot \mid s_0) = \prod_{t=0}^{T-1} \pi_T(\cdot \mid s_t)$ are the trajectory distributions induced by the student and the frozen teacher, respectively.

Since both policies are autoregressive and share the environment transitions, the tower property rewrites Eq. (1) as a sum of turn-level reverse KLs,

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{s_0 \sim \mathcal{D},\, \tau \sim P_\theta(\cdot|s_0)}\Big[\sum_{t=0}^{T-1} D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid s_t) \,\|\, \pi_T(\cdot \mid s_t)\big)\Big], \tag{2}$$

where the expectation is over the student's own trajectory, s_t is the state visited at turn t, and each summand is the reverse KL between the student and teacher response distributions at that state.

Applying the same decomposition one level down, to the tokens of each response $a_t = (y_1, \dots, y_{|a_t|})$, yields the exact token-level form of the OPD objective,

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{s_0 \sim \mathcal{D},\, \tau \sim P_\theta(\cdot|s_0)}\Big[\sum_{t=0}^{T-1} \sum_{\ell=1}^{|a_t|} D_{\mathrm{KL}}\big(p_\theta(\cdot \mid s_t, y_{<\ell}) \,\|\, p_T(\cdot \mid s_t, y_{<\ell})\big)\Big], \tag{3}$$

where ℓ indexes the tokens of the response, $y_{<\ell} = (y_1, \dots, y_{\ell-1})$ is the prefix preceding the ℓ-th token, and $p_\theta(\cdot \mid s_t, y_{<\ell})$ and $p_T(\cdot \mid s_t, y_{<\ell})$ are the student and teacher next-token distributions over the vocabulary V.

4 METHODOLOGY

4.1 TEACHER-STUDENT DIVERGENCE DRIFT

A turn is a decision about a whole response distribution, so we measure the importance of a turn by the divergence between the student's and the teacher's response distributions. Defined on the unit at which the agent acts rather than on its tokens, this measurement needs neither an extrinsic environment reward nor additional rollouts. We define the turn divergence as

$$D_{\text{turn}}(s) := D_{\text{KL}}\big(\pi_\theta(\cdot \mid s) \,\|\, \pi_T(\cdot \mid s)\big), \tag{4}$$

the turn-level reverse KL at state s, which is non-negative and vanishes only when the two response distributions coincide.

We further define the turn-level divergence score of a sampled response a at state s as

$$L(a \mid s) := \log \pi_\theta(a \mid s) - \log \pi_T(a \mid s). \tag{5}$$

We also define the policy-level teacher-student divergence as the trajectory-level reverse KL, which, using the autoregressive factorization, admits the exact turn-level decomposition of Eq. (2),

$$D_{\text{policy}}(\pi) := \mathbb{E}_{s_0\sim\mathcal{D}}\Big[D_{\text{KL}}\big(P_\theta(\cdot \mid s_0) \,\|\, P_T(\cdot \mid s_0)\big)\Big] = \mathbb{E}_{s_0\sim\mathcal{D},\, \tau\sim P_\theta(\cdot|s_0)}\Big[\sum_{t=0}^{T-1} D_{\text{turn}}(s_t)\Big]. \tag{6}$$

Measuring the turn's teacher-student divergence fixes what is measured, but not whether relatively high-divergence turns should be amplified. We now state the turn-level update exactly, then measure the drift it induces, which decides whether training on a response decreases the turn-level divergence.

Proposition 4.1.1 Turn-Level OPD Gradient. Fix a state s and treat the frozen teacher $\pi_T(\cdot \mid s)$ as independent of θ, with the student $\pi_\theta(\cdot \mid s)$ differentiable. The turn-level reverse KL satisfies

$$\nabla_\theta D_{\text{turn}}(s) = \mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\big[\big(\log \pi_\theta(a \mid s) - \log \pi_T(a \mid s)\big)\nabla_\theta \log \pi_\theta(a \mid s)\big]. \tag{7}$$

Hence, minimizing the turn-level reverse KL is equivalent to a REINFORCE-style ascent (Williams, 1992) whose scalar feedback is the negated divergence score, with a stop-gradient applied to it. We write it as

$$A^{\text{OPD}}(a, s) := \log \pi_T(a \mid s) - \log \pi_\theta(a \mid s) = -L(a \mid s). \tag{8}$$

Equivalently, the turn-level objective term

$$J_a(\pi) := A^{\text{OPD}}(a, s)\ \log \pi(a \mid s), \tag{9}$$

maximized with A^{OPD} held fixed, satisfies $\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\big[\nabla_\theta J_a(\pi_\theta)\big] = -\nabla_\theta D_{\text{turn}}(s)$. The proof is collected in Appendix A.1.

With the turn-level update pinned down, we measure how the teacher-student divergence changes along it with the Fisher-Rao metric on the open simplex $\Delta^\circ(\mathcal{A}_s)$ of the response distribution $\pi := \pi_\theta(\cdot \mid s)$ over the finite response set $\mathcal{A}_s$ at state s; the policy-simplex geometry and the local-KL interpretation of the metric are collected in Appendix A.2 (Amari & Nagaoka, 2000; Nielsen, 2020): the directional derivative of a functional along a tangent vector is the Fisher-Rao inner product $\langle\cdot,\cdot\rangle_{\text{Fisher-Rao}}$ with the functional's natural gradient grad_F (Kakade, 2001).

Theorem 4.1.2 Teacher-Student Divergence Drift under OPD. Fix a state s, with grad_F as above. The turn-level teacher-student divergence drift—the first-order drift of $D_{\text{turn}}(\pi)$ along the update direction $\text{grad}_F\, J_a(\pi)$—satisfies

$$\begin{aligned} D^{\text{turn}}_{\text{OPD}}(a; s) &:= \Big\langle \text{grad}_F\, D_{\text{turn}}(\pi),\ \text{grad}_F\, J_a(\pi)\Big\rangle_{\text{Fisher-Rao}} \\ &= A^{\text{OPD}}(a, s)\big(L(a \mid s) - D_{\text{turn}}(s)\big). \end{aligned} \tag{10}$$

Assume a local policy update under a frozen rollout distribution, that is, no gradients through P_θ when differentiating D_{policy}, taking the Fisher-Rao derivative with respect to the state-s marginal

alone. Then the policy-level teacher-student divergence drift induced by a sampled response a equals the visitation-weighted accumulation of the turn-level drift over the visits to s,

$$\begin{aligned} D_{\mathrm{OPD}}(a;s) &:= \Big\langle \mathrm{grad}_F\, D_{\mathrm{policy}}(\pi),\ \mathrm{grad}_F\, J_a(\pi) \Big\rangle_{\text{Fisher-Rao}} \\ &= \sum_{t=0}^{T-1} \mathbb{P}_{s_0\sim\mathcal{D},\,\tau\sim P_\theta(\cdot|s_0)}\big[s_t = s\big]\, A^{\mathrm{OPD}}(a,s)\big(L(a \mid s) - D_{\mathrm{turn}}(s)\big). \end{aligned} \tag{11}$$

where $\mathbb{P}_{s_0\sim\mathcal{D},\,\tau\sim P_\theta(\cdot|s_0)}[s_t = s]$ is the visit probability of state s at turn t. We define the turn-level relative divergence as

$$\rho(a \mid s) := L(a \mid s) - D_{\mathrm{turn}}(s). \tag{12}$$

The proof is in Appendix A.4.

Therefore, Theorem 4.1.2 shows that the turn-level teacher-student divergence drift induced by a sampled response is determined by the stop-gradient negated divergence score $A^{\mathrm{OPD}}(a,s)$ of the response and by the turn-level relative divergence $\rho(a \mid s)$. We calibrate the turn importance by the relative divergence of the response: a turn of larger relative divergence is more important and receives more of the distillation budget, and a turn of smaller relative divergence receives less. Notably the relative divergence is a property of the response distributions of the student and the teacher at the state rather than of the neural parameterization of either model, so turn importance is measured on the distributions themselves, and the measurement holds for any teacher-student pair. This is what leads to DT-OPD.

4.2 Group-Relative Divergence

Motivated by Theorem 4.1.2, we allocate the distillation budget of a prompt across its turns by the relative divergence of their responses: a turn of larger relative divergence receives a larger weight.

The relative divergence combines the turn's own divergence $D_{\mathrm{turn}}(s)$, an expectation over the response distribution, with the score $L(a \mid s)$ of a single sampled response. DT-OPD does not form the two terms separately. Instead, it compares a response's divergence with those of the other responses in its group, so that the relative magnitude within the group sets the turn's weight.

To reduce the sensitivity to the particular sampled response, we take the predictable part of the turn-level divergence score $L(a \mid s)$ from the Doob decomposition of Appendix A.3, which splits the score as $L(a \mid s) = \hat{L}(a \mid s) + M_{|a|}$, where $\hat{L}(a \mid s)$ is the predictable part, the sum of the token-level divergences,

$$\delta_\ell(a,s) := D_{\mathrm{KL}}\big(p_\theta(\cdot \mid s, y_{<\ell}) \,\|\, p_T(\cdot \mid s, y_{<\ell})\big), \tag{13}$$

the token-level term of Eq. (3), and $M_{|a|}$ is a zero-mean martingale; hence $\rho(a \mid s) = (\hat{L}(a \mid s) - D_{\mathrm{turn}}(s)) + M_{|a|}$, and $\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}[\hat{L}(a \mid s)] = D_{\mathrm{turn}}(s)$, the turn's own divergence. DT-OPD therefore evaluates only the predictable part, and neither L nor D_{turn} is computed explicitly: both are read off the truncated-support estimator of Eq. (13).

Length normalization. Here i indexes the rollouts generated for the same prompt and t the turn within a rollout, and $S_{i,t} := \{1, \ldots, |a_{i,t}|\}$ is the set of token positions of the (i,t)-th response. The predictable sum grows with response length, so we first normalize it by $|S_{i,t}|$,

$$\bar{D}_{i,t} = \frac{1}{|S_{i,t}|} \sum_{\ell \in S_{i,t}} \delta_\ell(a_{i,t}, s_{i,t}), \tag{14}$$

Group-relative normalization. Let G be the set of (i,t) pairs of all responses of all rollouts generated for the same prompt. We truncate outliers at preset group-level quantiles $q_{\mathrm{lo}} < q_{\mathrm{hi}}$ and min-max normalize the clipped scores,

$$\tilde{D}_{i,t} = \frac{\mathrm{clip}_{q_{\mathrm{lo}},q_{\mathrm{hi}}}(\bar{D}_{i,t}) - \min_{(j,n)\in G} \mathrm{clip}_{q_{\mathrm{lo}},q_{\mathrm{hi}}}(\bar{D}_{j,n})}{\max_{(j,n)\in G} \mathrm{clip}_{q_{\mathrm{lo}},q_{\mathrm{hi}}}(\bar{D}_{j,n}) - \min_{(j,n)\in G} \mathrm{clip}_{q_{\mathrm{lo}},q_{\mathrm{hi}}}(\bar{D}_{j,n}) + \varepsilon}, \tag{15}$$

where $\mathrm{clip}_{q_{\mathrm{lo}},q_{\mathrm{hi}}}$ clips each score $\bar{D}_{i,t}$ to the group's quantile range, and $\varepsilon > 0$ guards against division by zero. The map is invariant to a common shift of the scores $\bar{D}_{j,n}$ of the group and non-decreasing within the group (Lemma B.1.2), so the weights depend on the within-group order of the

scores rather than on the overall level of the group. Within a group, $\tilde{D}_{i,t}$ is a relative magnitude rather than an estimate of the relative divergence $\rho(a \mid s)$: the responses of different turns enter the same group, so a score is read against the group that produced it. The substitution is licensed at the level of the criterion rather than of exact values (Appendix B.1).

4.3 Divergence-Calibrated Turn Weights

How should a within-group relative score become a weight that preserves the ordering of the turns, without distorting the OPD objective? Because a relative score carries only the within-group ordering of the turns, a group-level change in how spread out the scores are would move every weight in the group. DT-OPD therefore exponentiates the score at a temperature $\lambda_D > 0$ and divides by its group mean, which holds the average weight of the group fixed:

$$w_{i,t}^{D} = \frac{\exp\big(+\lambda_D \, \tilde{D}_{i,t}\big)}{\frac{1}{|G|}\sum_{(j,n)\in G} \exp\big(+\lambda_D \, \tilde{D}_{j,n}\big) + \varepsilon}. \tag{16}$$

The resulting weights are strictly positive and increasing in $\tilde{D}$, so within each group a turn of higher relative score receives a larger weight than a turn of lower relative score. Their group-mean denominator keeps the group's average weight at approximately one, redistributing supervision within the group; the token-level budget is unchanged on average whenever the weights are uncorrelated with response length. The weights are computed from the current distributions and held fixed: the gradient is stopped at the weight, so it enters the objective as a constant multiplier and no gradient flows through it. Because the turn weights are derived from the teacher-student distributions that OPD already evaluates, they incur no additional inference cost: DT-OPD adds no extra rollouts and no additional evaluations of either model.

Broadcasting $w_{i,t}^{D}$ to every token of the (i,t)-th response turns it into a turn weight w_t, and plugging the stop-gradient weight w_t into every token-level term of Eq. (3) gives the DT-OPD objective:

$$\mathcal{L}_{\text{DT-OPD}}(\theta) = \mathbb{E}_{s_0\sim\mathcal{D},\, \tau\sim P_\theta(\cdot|s_0)}\left[\sum_{t=0}^{T-1}\sum_{\ell=1}^{|a_t|} w_t \cdot D_{\mathrm{KL}}\big(p_\theta(\cdot \mid s_t, y_{<\ell}) \,\big\|\, p_T(\cdot \mid s_t, y_{<\ell})\big)\right]. \tag{17}$$

The complete procedure is summarized in Algorithm 1 in Appendix B.2.

5 Experiments

5.1 Experimental Settings

Datasets and base settings. We evaluate on three multi-turn agentic environments: ALFWorld, a text-based embodied decision-making benchmark (Shridhar et al., 2020); Search-QA, a multi-turn retrieval-augmented question-answering benchmark (Jin et al., 2025); and WebShop, a web-interaction shopping benchmark (Yao et al., 2022a). We train students at three scales of the Qwen3 family (Yang et al., 2025a): Qwen3-0.6B and Qwen3-1.7B, distilled from the frozen Qwen3-4B GRPO teacher, and a Qwen3-4B student distilled from the frozen Qwen3-30B-A3B GRPO teacher. A further comparison with the Qwen2.5-3B student (Yang et al., 2025b) distilled from a frozen Qwen2.5-7B GRPO teacher is reported in Appendix D.1. For other detailed hyperparameter settings, please refer to Appendix C.

Comparative algorithm. We compare DT-OPD with the untrained student base, GRPO (Shao et al., 2024), OPD (Agarwal et al., 2024), and the turn-aware distillation methods SOD (Zhong et al., 2026) and TCOD (Wang et al., 2026). GRPO and SOD are trained with environment rewards, whereas OPD, TCOD, and DT-OPD learn from the teacher distribution alone. Baseline descriptions are provided in Appendix C.2.

5.2 Performance Comparison

Performance on ALFWorld and Search-QA. To compare DT-OPD with baselines on ALFWorld and Search-QA, we use two GRPO-trained teachers: Qwen3-4B for the Qwen3-0.6B and Qwen3-1.7B students, and Qwen3-30B-A3B for the Qwen3-4B student. The performance comparison is

Table 1: Main results on ALFWorld and Search-QA. We report success rate (SR, %). Vanilla is the untrained base model, and the other methods distill from a frozen teacher. The suffix -G marks a GRPO-trained model. Best results are in **bold**, second best underlined.

Method	ALFWorld							Search-QA							
	Pick	Look	Clean	Heat	Cool	Pick2	Avg.	NQ	Triv	Pop	Hotp	2Wk	MuS	Bam	Avg.
Qwen3-4B-G	89.2	58.8	47.6	84.6	71.4	73.3	72.7	45.3	63.0	47.4	41.9	40.9	15.6	70.2	47.0
	Qwen3-0.6B Student (Teacher: Qwen3-4B-G)														
Vanilla	0.0	0.0	0.0	0.0	4.8	0.0	0.8	11.4	16.3	13.0	10.5	21.0	1.3	1.2	14.1
GRPO	25.9	25.0	40.0	40.0	16.7	4.5	25.8	**42.9**	51.1	44.1	31.7	30.9	7.4	33.3	38.8
OPD	80.7	54.6	73.1	86.7	70.6	**64.3**	72.7	38.0	54.1	45.4	34.1	29.1	9.9	64.1	40.0
SOD	**90.3**	45.5	80.8	93.3	58.8	**64.3**	75.0	38.8	53.4	45.5	34.0	28.3	9.8	15.3	38.7
TCOD	83.9	**63.6**	73.1	**100.0**	70.6	**64.3**	75.8	40.9	53.1	39.3	30.0	**32.6**	6.6	33.3	37.9
DT-OPD	88.9	62.5	**86.7**	83.3	**77.8**	63.6	**78.9**	39.2	**54.7**	**45.5**	**34.4**	31.4	**10.6**	**64.1**	**40.8**
	Qwen3-1.7B Student (Teacher: Qwen3-4B-G)														
Vanilla	3.7	0.0	0.0	0.0	0.0	7.7	2.3	29.7	46.9	36.9	23.5	19.4	6.5	9.7	30.2
GRPO	71.7	41.7	36.4	40.0	31.8	31.6	46.1	40.0	**58.9**	43.5	35.4	30.3	12.0	65.7	40.8
OPD	74.2	**72.7**	80.8	93.3	**82.4**	60.7	75.8	42.3	57.9	46.0	37.4	35.0	14.0	69.8	43.4
SOD	80.7	54.6	76.9	80.0	70.6	57.1	71.1	42.0	58.2	45.5	38.1	33.9	14.3	**70.6**	43.0
TCOD	77.4	54.6	84.6	93.3	58.8	64.3	73.4	39.8	57.1	42.5	36.5	**38.6**	12.3	33.3	42.2
DT-OPD	**87.1**	63.6	**84.6**	**93.3**	64.7	**67.9**	**78.1**	**43.7**	58.7	**46.8**	**38.2**	35.6	**14.9**	69.4	**44.0**
Qwen3-30B-A3B-G	75.0	66.7	100.0	83.3	90.5	72.7	80.5	47.4	65.7	45.4	44.7	47.7	21.9	66.7	49.3
	Qwen3-4B Student (Teacher: Qwen3-30B-A3B-G)														
Vanilla	55.6	33.3	7.1	14.3	4.8	19.2	22.7	31.6	47.7	29.4	23.9	30.0	7.8	33.3	31.7
GRPO	89.2	58.8	47.6	**84.6**	71.4	73.3	72.7	45.3	63.0	**47.4**	41.9	40.9	15.6	70.2	47.0
OPD	**93.3**	63.6	80.8	33.3	76.9	78.3	78.1	**48.1**	62.7	45.4	41.8	**44.9**	**19.0**	66.7	47.5
SOD	80.8	58.3	**85.7**	68.8	72.7	**83.9**	77.3	42.0	55.1	39.1	38.2	39.9	17.4	**100.0**	45.9
TCOD	84.8	**75.0**	80.8	58.3	76.9	69.6	76.6	40.9	58.0	40.6	37.4	41.2	16.9	71.4	47.0
DT-OPD	92.6	68.8	83.3	80.0	**77.8**	63.6	**78.9**	46.2	**64.5**	46.0	**43.4**	44.4	18.6	73.0	**48.1**

summarized in Table 1. DT-OPD achieves the best overall performance across both benchmarks and all three student sizes. Across the three students it attains 78.9%, 78.1%, and 78.9% on ALFWorld and 40.8%, 44.0%, and 48.1% on Search-QA, and the margins over the strongest same-setting baseline are +3.1, +2.3, and +0.8 on ALFWorld and +0.8, +0.6, and +0.6 on Search-QA. This shows that calibrating the turn importance by the relative divergence of a response is effective. We also verify DT-OPD on the Qwen2.5 family (see Appendix D.1).

Table 2: WebShop results. We report Score and Acc (%) for the Qwen3-0.6B and Qwen2.5-3B students, distilled from the Qwen3-4B GRPO and Qwen2.5-7B GRPO teachers, respectively. Vanilla is the untrained base model. Best results are in **bold**, second best underlined.

Metric	Vanilla	GRPO	OPD	SOD	TCOD	DT-OPD	Teacher
	Qwen3-0.6B Student (Teacher: Qwen3-4B GRPO)						
Score	33.1	44.9	88.6	90.6	87.0	**91.0**	87.4
Acc	2.3	16.4	**82.0**	78.8	79.2	**82.0**	79.7
	Qwen2.5-3B Student (Teacher: Qwen2.5-7B GRPO)						
Score	6.5	79.8	81.1	82.2	82.4	**84.7**	90.7
Acc	0.8	63.3	69.8	71.4	70.8	**75.8**	79.7

Performance on WebShop. To verify that DT-OPD is also effective on other benchmarks, we further compare it with the same baselines on WebShop; Table 2 reports both metrics for the Qwen3-0.6B and Qwen2.5-3B students. On the Qwen3-0.6B student DT-OPD attains the best Score of the block, 91.0%, 0.4 points over SOD, and matches its best Acc at 82.0%; on the Qwen2.5-3B student

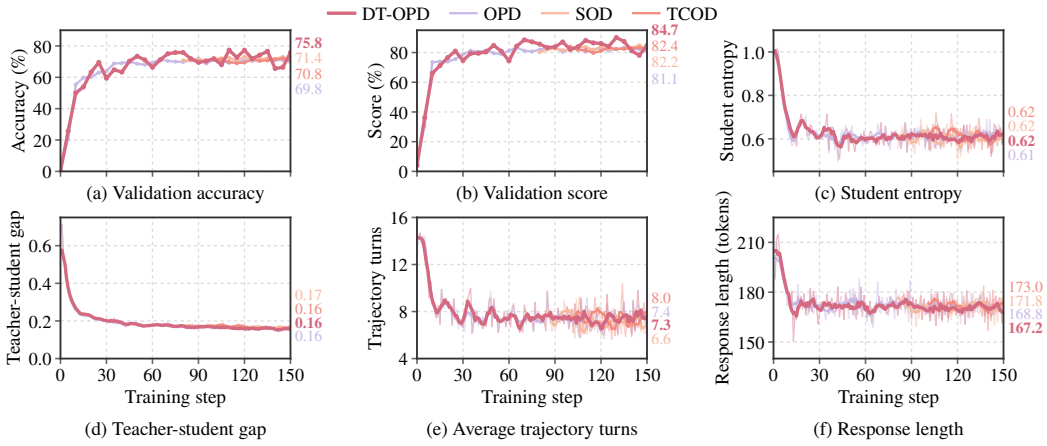


Figure 1: Training dynamics on WebShop with the Qwen2.5-3B student (Teacher: Qwen2.5-7B GRPO). (a) Validation accuracy. (b) Validation score. (c) Student entropy. (d) Teacher-student gap. (e) Average trajectory turns. (f) Response length.

it attains both, 84.7% and 75.8%, 2.3 and 4.4 points above the strongest baselines. This shows that DT-OPD is also effective in a multi-turn interactive online shopping environment.

Training dynamics. Figure 1 reports the WebShop training dynamics of DT-OPD, OPD, SOD, and TCOD with the Qwen2.5-3B student and its frozen Qwen2.5-7B GRPO teacher. In panels (a) and (b), DT-OPD achieves the highest validation accuracy, 75.8% against 69.8%, 71.4%, and 70.8% for OPD, SOD, and TCOD, and the highest validation score, 84.7% against 81.1%, 82.2%, and 82.4%. Panels (c)–(f) show that the higher success is not bought by longer interactions or closer imitation: all four runs end in the same band on trajectory turns (6.6–8.0 per episode), response length (167–173 tokens), student entropy (0.61–0.62), and the teacher-student gap (0.16–0.17).

5.3 Ablation Experiment

To better understand where the gain of DT-OPD comes from, we conduct controlled ablation studies on ALFWorld with the Qwen3-1.7B student and the frozen Qwen3-4B GRPO teacher. The only difference lies in how the weights w^D of Eq. (16) are constructed or applied. We compare the following variants: **OPD**, the vanilla distillation objective without divergence-calibrated turn weighting; **DT-OPD**, our full method of Section 4; **DT-OPD**$_{\text{shuffle}}$, which computes w^D as in DT-OPD but randomly permutes the weights within each group, keeping their distribution while destroying the turn–weight alignment; **DT-OPD**$_{\text{reverse}}$, which negates the temperature λ_D, so that low-divergence turns are amplified and high-divergence turns are suppressed; and **DT-OPD**$_{\text{traj-norm}}$, which replaces the group-relative normalization of Eq. (15) with a trajectory-level one, rescaling each score against its own rollout instead of the group.

Table 3: Ablation of the weighting scheme on ALFWorld with the Qwen3-1.7B student and the frozen Qwen3-4B GRPO teacher, under the same protocol as the main experiments. Best results are in **bold**, second best underlined.

Variant	ALFWorld						
	Pick	Look	Clean	Heat	Cool	Pick2	Avg.
OPD	74.2	**72.7**	80.8	93.3	**82.4**	60.7	<u>75.8</u>
DT-OPD$_{\text{shuffle}}$	<u>86.2</u>	46.7	**85.2**	<u>90.9</u>	<u>80.8</u>	50.0	75.0
DT-OPD$_{\text{reverse}}$	83.9	54.6	80.8	80.0	76.5	<u>64.3</u>	75.0
DT-OPD$_{\text{traj-norm}}$	79.3	46.7	**85.2**	<u>90.9</u>	63.6	55.0	70.3
DT-OPD	**87.1**	<u>63.6</u>	<u>84.6</u>	**93.3**	64.7	**67.9**	**78.1**

The ablation results are reported in Table 3: the full DT-OPD attains the best ALFWorld average (78.1%), confirming that the gain comes from ordering the weights by the divergence rather than from non-uniform weighting alone. DT-OPD$_{\mathrm{shuffle}}$ drops to 75.0%, below the OPD baseline (75.8%): the benefit is not the fine-grained turn-level rescaling itself, but the alignment of each divergence signal with its own turn. DT-OPD$_{\mathrm{reverse}}$ likewise drops to 75.0%: the reversed map suppresses the high-divergence turns, for which the drift of Theorem 4.1.2 is negative so that training decreases the turn-level divergence, and amplifies the turns the student already reproduces, so the divergence-decreasing pressure of those turns is removed and the weights must follow the divergence—not reverse it. DT-OPD$_{\mathrm{traj-norm}}$ falls further, to 70.3%: normalizing within a single rollout draws its reference set from far fewer responses, so the scores inherit sampling noise that the group-relative proxy of Eq. (15) avoids.

5.4 Computational Cost

DT-OPD calibrates the turn importance by the divergence of each response, which depends directly on logits that are already computed, so, relative to the vanilla OPD run, which consumes the same rollout budget and follows the same evaluation rule (Appendix C), it introduces no extra rollouts and no additional model to train. We quantify this advantage with two cost measures: the stage-by-stage per-step latency, and the training and validation time across student scales. Table 4 reports the first measure for the Qwen3-0.6B student, alongside OPD, SOD, and TCOD.

Table 4: Per-step latency decomposition on ALFWorld with the Qwen3-0.6B student, in seconds. For each training stage, the table reports the latency of DT-OPD, OPD, SOD, and TCOD, together with their total per-step latency.

Stage	DT-OPD	OPD	SOD	TCOD
Rollout generation	182.1	198.6	189.7	196.5
Old-policy log-probabilities	6.6	5.1	5.1	5.6
Teacher log-probabilities	17.8	16.8	16.7	18.0
Weight construction + advantage	0.54	0.45	0.55	0.41
Policy update	19.1	19.6	19.2	20.8
Total per step	419.1	431.7	418.6	432.0

Overall, the divergence-calibrated weight construction adds only 0.09 s to the advantage stage, about 0.02% of the step, and the per-step total matches the OPD run ($0.97\times$ in Table 4); the remaining stage-level differences reflect run-to-run variation rather than the weight construction. The second measure, the training and validation time across student scales, is reported in Table 8 of Appendix D.

6 Conclusion

We proposed Divergence-Calibrated Turn Importance for Multi-Turn Agentic On-Policy Distillation (DT-OPD) to answer how supervision should be allocated across turns by their importance when no environment reward is available. We demonstrated that, under natural-gradient updates, the teacher-student divergence drift induced by a sampled response is determined by the response's stop-gradient negated divergence score and by its relative divergence. Our analysis lifts the divergence from the token level to the turn level, aligning it with the unit at which the agent acts. DT-OPD therefore turns the drift into a practical turn-level teacher-student divergence proxy and redistributes supervision across turns accordingly. Extensive experiments on ALFWorld, Search-QA, and WebShop, distilling Qwen3 and Qwen2.5 students from 0.6B to 4B using GRPO teachers up to Qwen3-30B-A3B, show that DT-OPD achieves the best overall performance across the three benchmarks and all student scales, and ablation studies and training dynamics further confirm that both of our core contributions are essential to this performance. DT-OPD adds no extra rollouts and no trainable parameters.

AI USE STATEMENT

Generative AI tools are used as research and writing assistants during the development of this work. Specifically, they are used to assist with checking and reorganizing mathematical derivations, discussing experimental designs and interpretations, identifying potentially relevant literature, improving the clarity and structure of the manuscript, and assisting with the design of scientific figures. All AI-assisted technical content is independently reviewed and verified by the authors. The authors check mathematical derivations, manually verify cited literature against the original sources, and validate all reported experimental results and conclusions against the authors' own experiments. The authors take full responsibility for the final content of the paper, including all claims, equations, figures, and experimental results.

ETHICS STATEMENT

This work is a training-method study evaluated on public simulated benchmarks (ALFWorld, Search-QA, and WebShop) with publicly released base models. It does not involve human subjects, proprietary or sensitive data, or experiments that raise privacy, security, or legal concerns, and no institutional review was required. The method only reallocates the teacher–student distillation signal that the base objective already computes across turns: it adds no environment reward, no extra sampling, and no extra model inference, and it leaves the interface through which the trained policy acts unchanged. Our intent is to promote transparency, safety, and responsible analysis of language models, and we recommend that future work consider these implications in line with the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

We have taken steps to ensure the reproducibility of our results. All definitions, assumptions, and theoretical proofs are included in the main text and appendix: Appendix A.1 proves Proposition 4.1.1 and Appendix A.4 proves Theorem 4.1.2. Detailed algorithms (Algorithm 1) and mathematical derivations are provided in Appendix B.2 and Appendix B.1. Experimental procedures, benchmarks, model families used (the Qwen3 and Qwen2.5 families), and metrics are described in Section 5 and Appendix C. We will open-source, upon acceptance, the code implementing DT-OPD, constructing its divergence-calibrated turn weights, and reproducing every reported result, ensuring that all empirical results reported can be replicated.

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 and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2000.

Chishui Chen, Yaoyou Fan, Te Sun, Yi Yang, Chenghao Sun, Delin Mao, Hongbo Qiao, Zuowei Zhang, Junxi Wang, Chenxing Sun, et al. Look ahead before you distill: Future trajectory validation of teacher guidance for agentic on-policy distillation. *arXiv preprint arXiv:2608.01953*, 2026a.

Mingguang Chen, Licheng Wang, and Bo Qu. Recursive self-improvement in AI: From bounded self-refinement to autonomous research loops. *arXiv preprint arXiv:2607.07663*, 2026b.

Xianwei Chen, Shimin Zhang, and Jibin Wu. f-OPD: Stabilizing long-horizon on-policy distillation with freshness-aware control. *arXiv preprint arXiv:2605.17862*, 2026c.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2Web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36:28091–28114, 2023.

Guanting Dong, Xiaoshuai Song, Yuyang Hu, Jiajie Jin, Chenghao Zhang, Yifei Chen, Xiaoxi Li, Huaying Yuan, Xinyu Yang, Tongyu Wen, Jiejun Tan, Hongjin Qian, Shijue Huang, Junting Lu, Zhenyu Li, Wanjun Zhong, Yutao Zhu, Tat-Seng Chua, Zhicheng Dou, and Ji-Rong Wen. Towards long-horizon agents: A survey. *Preprints*, 2026. doi: 10.20944/preprints202607.1328.v1.

Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, et al. Agent AI: Surveying the horizons of multimodal interaction. *arXiv preprint arXiv:2401.03568*, 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.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, 2020.

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.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017.

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

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th ACM Symposium on Operating Systems Principles (SOSP)*, pp. 611–626, 2023.

Gengsheng Li, Mao Zheng, Mingyang Song, Ruiqi Liu, Tianyu Yang, Jie Sun, Qiyong Zhong, Haiyun Guo, Junfeng Fang, Dan Zhang, et al. On-policy distillation with curriculum turn-level guidance for multi-turn agents. *arXiv preprint arXiv:2606.15912*, 2026a.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, et al. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 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.

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.

Zhengxi Lu, Zhiyuan Yao, Zhuowen Han, Zi-Han Wang, Jinyang Wu, Qi Gu, Xunliang Cai, Weiming Lu, Jun Xiao, Yueting Zhuang, et al. Self-distilled agentic reinforcement learning. *arXiv preprint arXiv:2605.15155*, 2026.

Junyu Luo, Weizhi Zhang, Ye Yuan, Yusheng Zhao, Junwei Yang, Yiyang Gu, Bohan Wu, Binqi Chen, Ziyue Qiao, Qingqing Long, et al. Large language model agent: A survey on methodology, applications and challenges. *arXiv preprint arXiv:2503.21460*, 2025.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 9802–9822, 2023.

Amir M Mansourian, Rozhan Ahmadi, Masoud Ghafouri, Amir Mohammad Babaei, Elaheh Badali Golezani, Zeynab Yasamani Ghamchi, Vida Ramezanian, Alireza Taherian, Kimia Dinashi, Amirali Miri, et al. A comprehensive survey on knowledge distillation. *arXiv preprint arXiv:2503.12067*, 2025.

Frank Nielsen. An elementary introduction to information geometry. *Entropy*, 22(10):1100, 2020.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711, 2023.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551, 2023.

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.

Mingyang Song and Mao Zheng. A survey of on-policy distillation for large language models. *arXiv preprint arXiv:2604.00626*, 2026.

Changhui Sun, Lanbo Liu, Hang Lei, Tong Ling, Jiahang Xie, Zhiyong Zheng, Yujia Wang, Hao Liu, Feng Xiao, Lu Liu, et al. Step-level on-policy distillation: Interpolating between on-policy distillation and supervised fine-tuning. *arXiv preprint arXiv:2608.16333*, 2026.

Qitai Tan, Zefang Zong, Yang Li, and Peng Chen. ATOD: Annealed turn-aware on-policy distillation for multi-turn autonomous agents. *arXiv preprint arXiv:2606.27814*, 2026.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y Charles, et al. Kimi K3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

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

Quan Wei, Siliang Zeng, Chenliang Li, Zhongruo Wang, William Brown, Oana Frunza, Wei Deng, Anderson Schneider, Yuriy Nevmyvaka, Yang Zhao, et al. Reinforcing multi-turn reasoning in LLM agents via fine-grained reward structure and credit assignment. *arXiv preprint arXiv:2505.11821*, 2025.

Muning Wen, Ziyu Wan, Weinan Zhang, Jun Wang, and Ying Wen. Reinforcing language agents via policy optimization with action decomposition. *arXiv preprint arXiv:2405.15821*, 2024.

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

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *Science China information sciences*, 68(2):121101, 2025.

Yuchen Xia, Qianguo Sun, Chao Song, Junlong Wu, Yiyan Qi, and Yunjian Xu. DASH-OPD: Discrepancy-aware switching with hysteresis for on-policy distillation. *arXiv preprint arXiv:2607.29078*, 2026.

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

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025b.

John Yang, Carlos Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. SWE-agent: Agent-computer interfaces enable automated software engineering. *Advances in Neural Information Processing Systems*, 37:50528–50652, 2024.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2369–2380, 2018.

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, 2022a.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022b.

Zhaochen Yu, Yingcheng Wu, Zhenfei Yin, Kaiyuan Chen, Zhe Zhao, Mengdi Wang, Shuicheng Yan, and Ling Yang. Recursive experiential-working memory evolution for long-horizon agent harnesses. *arXiv preprint arXiv:2608.24876*, 2026.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. GLM-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 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.

Yuhang Zhou, Lizhu Zhang, Yifan Wu, Mingyi Wang, Bo Peng, Jiayi Liu, Xiangjun Fan, and Zhuokai Zhao. SAGE-OPD: Selective agent-guided intervention for multi-turn on-policy distillation. *arXiv preprint arXiv:2606.19659*, 2026a.

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

A Theoretical Proofs

A.1 Proof of Proposition 4.1.1

Proof. Fix a state s with finite response set $\mathcal{A}_s$, and treat the frozen teacher $\pi_T(\cdot \mid s)$ as independent of θ.

Step 1: the reverse-KL gradient. Write the turn-level reverse KL of Eq. (4) as a sum over the finite response set,

$$D_{\text{turn}}(s) = \sum_{a \in \mathcal{A}_s} \pi_\theta(a \mid s)\big(\log \pi_\theta(a \mid s) - \log \pi_T(a \mid s)\big). \tag{18}$$

Taking the gradient gives

$$\begin{aligned}
\nabla_\theta D_{\text{turn}}(s) &= \sum_{a \in \mathcal{A}_s} \big(\log \pi_\theta(a \mid s) - \log \pi_T(a \mid s)\big)\nabla_\theta \pi_\theta(a \mid s) + \sum_{a \in \mathcal{A}_s} \pi_\theta(a \mid s)\, \nabla_\theta \log \pi_\theta(a \mid s) \\
&= \sum_{a \in \mathcal{A}_s} \Big(\log \frac{\pi_\theta(a \mid s)}{\pi_T(a \mid s)} + 1\Big)\nabla_\theta \pi_\theta(a \mid s) \\
&= \sum_{a \in \mathcal{A}_s} \log \frac{\pi_\theta(a \mid s)}{\pi_T(a \mid s)}\, \nabla_\theta \pi_\theta(a \mid s), \qquad \sum_{a \in \mathcal{A}_s} \nabla_\theta \pi_\theta(a \mid s) = \nabla_\theta 1 = 0, \\
&= \sum_{a \in \mathcal{A}_s} \pi_\theta(a \mid s) \log \frac{\pi_\theta(a \mid s)}{\pi_T(a \mid s)}\, \nabla_\theta \log \pi_\theta(a \mid s) \\
&= \mathbb{E}_{a \sim \pi_\theta(\cdot \mid s)}\big[L(a \mid s)\, \nabla_\theta \log \pi_\theta(a \mid s)\big],
\end{aligned} \tag{19}$$

where the first line applies the product rule, the second merges the two terms using $\pi_\theta(a \mid s)\nabla_\theta \log \pi_\theta(a \mid s) = \nabla_\theta \pi_\theta(a \mid s)$, and the third drops the $+1$ term, which vanishes after summing over a since $\sum_{a \in \mathcal{A}_s} \nabla_\theta \pi_\theta(a \mid s) = \nabla_\theta 1 = 0$; the fourth applies the score-function identity $\nabla_\theta \pi_\theta(a \mid s) = \pi_\theta(a \mid s)\nabla_\theta \log \pi_\theta(a \mid s)$. Therefore the negative gradient used for minimization is the REINFORCE ascent estimator with the stop-gradient negated divergence score $A^{\text{OPD}}(a, s) = \log \pi_T(a \mid s) - \log \pi_\theta(a \mid s)$,

$$\begin{aligned}
-\nabla_\theta D_{\text{turn}}(s) &= \mathbb{E}_{a \sim \pi_\theta(\cdot \mid s)}\big[\big(\log \pi_T(a \mid s) - \log \pi_\theta(a \mid s)\big)\nabla_\theta \log \pi_\theta(a \mid s)\big] \\
&= \mathbb{E}_{a \sim \pi_\theta(\cdot \mid s)}\big[A^{\text{OPD}}(a, s)\, \nabla_\theta \log \pi_\theta(a \mid s)\big],
\end{aligned} \tag{20}$$

which is the ascent announced in the proposition: the stop-gradient negated divergence score $A^{\text{OPD}}(a, s)$ of Eq. (8), evaluated at the rollout, enters the update only as scalar feedback. Since it is a scalar, $\nabla_\theta J_a(\pi_\theta) = A^{\text{OPD}}(a, s)\, \nabla_\theta \log \pi_\theta(a \mid s)$, so Eq. (20) reads $\mathbb{E}_{a \sim \pi_\theta(\cdot \mid s)}[\nabla_\theta J_a(\pi_\theta)] = -\nabla_\theta D_{\text{turn}}(s)$, the turn-level objective term of Eq. (9).

Step 2: the closing identities. Immediate from Eqs. (5) and (8), $A^{\text{OPD}}(a, s)$ is the negated divergence score,

$$A^{\text{OPD}}(a, s) = \log \pi_T(a \mid s) - \log \pi_\theta(a \mid s) = -\big(\log \pi_\theta(a \mid s) - \log \pi_T(a \mid s)\big) = -L(a \mid s),$$

and taking the expectation under $a \sim \pi_\theta(\cdot \mid s)$ recovers the turn-level divergence of Eq. (4),

$$\mathbb{E}_{a \sim \pi_\theta(\cdot \mid s)}\big[L(a \mid s)\big] = \sum_{a \in \mathcal{A}_s} \pi_\theta(a \mid s) \log \frac{\pi_\theta(a \mid s)}{\pi_T(a \mid s)} = D_{\text{turn}}(s).$$

This proves Proposition 4.1.1.

A.2 Policy Simplex

For a fixed state s with finite action space $|\mathcal{A}_s| = m$, the policy $\pi = \big(\pi(a \mid s)\big)_{a \in \mathcal{A}_s}$ lies on the open simplex

$$\Delta^\circ(\mathcal{A}_s) := \Big\{\pi \in \mathbb{R}^m : \sum_{a \in \mathcal{A}_s} \pi(a \mid s) = 1 \ \text{ and } \ \pi(a \mid s) > 0 \ \text{ for all } a \in \mathcal{A}_s\Big\}, \tag{21}$$

equipped with the Fisher-Rao metric

$$g_\pi(u, v) := \sum_{a=1}^{m} \frac{u_a\, v_a}{\pi(a \mid s)}, \tag{22}$$

whose tangent space at π is the set of zero-sum vectors,

$$T_\pi \Delta^\circ(\mathcal{A}_s) = \big\{x \in \mathbb{R}^m : \mathbf{1}^\top x = 0\big\}. \tag{23}$$

The Fisher-Rao metric is the infinitesimal quadratic form induced by the KL divergence: for any tangent perturbation $\delta \in T_\pi \Delta^\circ(\mathcal{A}_s)$, i.e. $\sum_{a \in \mathcal{A}_s} \delta_a = 0$,

$$D_{\mathrm{KL}}(\pi + \delta \,\|\, \pi) = \frac{1}{2} \sum_{a \in \mathcal{A}_s} \frac{\delta_a^2}{\pi(a \mid s)} + o\big(\|\delta\|^2\big) = \frac{1}{2}\, g_\pi(\delta, \delta) + o\big(\|\delta\|^2\big). \tag{24}$$

Thus the Fisher-Rao metric measures the local size of a policy update in the same units as a local KL trust region. This subsection collects the simplex background used by the natural-gradient computations of Theorem 4.1.2 (proof in Appendix A.4); see the textbook treatment in (Nielsen, 2020).

A.3 Doob Decomposition of the Divergence Score

Proposition A.3.1 Doob Decomposition of the Divergence Score. Fix a state s and let $a = (y_1, \ldots, y_{|a|}) \sim \pi_\theta(\cdot \mid s)$ be a sampled response, written with the prefix notation $y_{<\ell} = (y_1, \ldots, y_{\ell-1})$ of Eq. (3) and with realized length $|a| \le n$ bounded by the maximum output length n of Section 3.1. Let $(\mathcal{F}_\ell)_{\ell \ge 0}$ be the autoregressive filtration, in which $\mathcal{F}_\ell$ is the information carried by the first ℓ tokens, and define the realized per-position log-ratio

$$X_\ell(a, s) := \log \frac{p_\theta(y_\ell \mid s, y_{<\ell})}{p_T(y_\ell \mid s, y_{<\ell})},$$

so that

$$L(a \mid s) = \sum_{\ell=1}^{|a|} X_\ell(a, s) = \sum_{\ell=1}^{|a|} \delta_\ell(a, s) + M_{|a|},$$

where $\delta_\ell(a, s) := D_{\mathrm{KL}}\big(p_\theta(\cdot \mid s, y_{<\ell}) \,\|\, p_T(\cdot \mid s, y_{<\ell})\big)$ is the token-level divergence of Eq. (3) and $M_k := \sum_{\ell \le k} \big(X_\ell(a, s) - \delta_\ell(a, s)\big)$ is a martingale with respect to $(\mathcal{F}_k)$ that satisfies $\mathbb{E}[M_k] = 0$ at every k; stopped at the realized length, $M_{|a|}$ is zero-mean as well.

Proof. The state s is fixed throughout, and the teacher is assumed to charge every token the student can sample, $p_T(y \mid s, y_{<\ell}) > 0$ for each y in the support $V_\ell \subseteq V$ of $p_\theta(\cdot \mid s, y_{<\ell})$; this regularity condition is the one that makes the KL of Eq. (3) finite, and it keeps every realized log-ratio $X_\ell(a, s)$ finite. The prefix $y_{<\ell}$ is a function of the first $\ell - 1$ tokens, so it determines the pair of next-token distributions, and with them the conditional mean $\delta_\ell(a, s)$.

Step 1: the conditional mean of the realized log-ratio. Condition on the prefix, that is, on $\mathcal{F}_{\ell-1}$: the ℓ-th token is drawn from the student's own next-token distribution, $y_\ell \sim p_\theta(\cdot \mid s, y_{<\ell})$, and the prefix is no longer random, so the conditional expectation of the realized log-ratio is

$$\mathbb{E}\big[X_\ell(a, s) \mid \mathcal{F}_{\ell-1}\big] = \sum_{y \in V_\ell} p_\theta(y \mid s, y_{<\ell}) \log \frac{p_\theta(y \mid s, y_{<\ell})}{p_T(y \mid s, y_{<\ell})} = \delta_\ell(a, s), \tag{25}$$

the sum being the reverse KL between the two next-token distributions at the realized prefix. The conditional mean is thus a function of the prefix alone; it does not depend on the token y_ℓ that is about to be drawn, which is what makes the first sum of the decomposition predictable.

Step 2: martingale differences. Since $\delta_\ell(a, s)$ is $\mathcal{F}_{\ell-1}$-measurable by the previous step, subtracting it from both sides of Eq. (25) leaves a centered increment,

$$\mathbb{E}\big[X_\ell(a, s) - \delta_\ell(a, s) \mid \mathcal{F}_{\ell-1}\big] = \mathbb{E}\big[X_\ell(a, s) \mid \mathcal{F}_{\ell-1}\big] - \delta_\ell(a, s) = 0, \tag{26}$$

so the centered increments are martingale differences. Each $X_\ell(a,s)$ is a function of the first ℓ tokens and each $\delta_\ell(a,s)$ a function of the first $\ell-1$, so the partial sums $M_k = \sum_{\ell=1}^{k}(X_\ell(a,s)-\delta_\ell(a,s))$, with the empty sum $M_0 := 0$, are $\mathcal{F}_k$-measurable with finite first moment, and Eq. (26) with the linearity of the conditional expectation gives

$$\mathbb{E}\big[M_k \mid \mathcal{F}_{k-1}\big] = M_{k-1},$$

that is, (M_k) is a martingale with respect to $(\mathcal{F}_k)$; in particular $\mathbb{E}[M_k] = \mathbb{E}[M_0] = 0$ at every k, by the tower property of the conditional expectation.

Step 3: summation at the realized length. The response distribution factorizes autoregressively, so the score of Eq. (5) splits token by token,

$$L(a \mid s) = \log \frac{\pi_\theta(a \mid s)}{\pi_T(a \mid s)} = \sum_{\ell=1}^{|a|} \log \frac{p_\theta(y_\ell \mid s, y_{<\ell})}{p_T(y_\ell \mid s, y_{<\ell})} = \sum_{\ell=1}^{|a|} X_\ell(a,s),$$

and inserting the conditional mean at every position splits that sum into its predictable part and the martingale,

$$L(a \mid s) = \sum_{\ell=1}^{|a|} \Big(\delta_\ell(a,s) + \big(X_\ell(a,s) - \delta_\ell(a,s)\big) \Big) = \sum_{\ell=1}^{|a|} \delta_\ell(a,s) + M_{|a|},$$

which is the decomposition stated in the proposition.

Step 4: zero mean at the realized length. The realized length $\tau := |a|$ is a stopping time of $(\mathcal{F}_\ell)$: the response ends as soon as its own tokens end it or the cap n is reached, so the event $\{\tau \le k\}$ is decided by the first k tokens in either case and belongs to $\mathcal{F}_k$, while $\tau \le n$ makes the stopping time bounded. Writing the stopped martingale as the sum of its increments, each entering only while the response is still running,

$$M_{|a|} = \sum_{k=1}^{|a|} \big(M_k - M_{k-1}\big) = \sum_{k=1}^{n} \big(X_k(a,s) - \delta_k(a,s)\big) \mathbf{1}\{k \le \tau\},$$

the running event is the complement of $\{\tau \le k-1\}$, hence $\{k \le \tau\} \in \mathcal{F}_{k-1}$ and it is legitimate to pull it out of the conditional expectation,

$$\mathbb{E}\big[\big(X_k(a,s) - \delta_k(a,s)\big)\mathbf{1}\{k \le \tau\}\big] = \mathbb{E}\big[\mathbf{1}\{k \le \tau\}\, \mathbb{E}\big[X_k(a,s) - \delta_k(a,s) \mid \mathcal{F}_{k-1}\big]\big] = 0,$$

with $\mathbf{1}\{\cdot\}$ the indicator of an event and with the increment at position k evaluated only on the event that the response has not ended earlier; summing over k gives $\mathbb{E}[M_{|a|}] = 0$. Stopping at the realized length therefore preserves the zero mean: the stopped value is centered even though the length of the response is not fixed in advance but sampled together with the tokens.

Step 5: the score, its mean, and its centered form. Taking expectations in Step 3 and using $\mathbb{E}[M_{|a|}] = 0$,

$$\mathbb{E}\big[L(a \mid s)\big] = \mathbb{E}\Big[\sum_{\ell=1}^{|a|} \delta_\ell(a,s)\Big],$$

so the predictable sum agrees with the score in mean and differs from it, response by response, by the martingale; combined with $\mathbb{E}[L(a \mid s)] = D_{\mathrm{turn}}(s)$ of Eq. (4), this is the path-sum identity of Eq. (3) read at the level of a single response. Subtracting the state-level constant $D_{\mathrm{turn}}(s)$ from both sides of the decomposition, the centered form used by the analysis of Appendix B.1 follows,

$$L(a \mid s) - D_{\mathrm{turn}}(s) = \Big(\sum_{\ell=1}^{|a|} \delta_\ell(a,s) - D_{\mathrm{turn}}(s)\Big) + M_{|a|},$$

which splits the relative divergence that drives the weighting into a predictable part and a zero-mean martingale. This completes the proof of Proposition A.3.1.

A.4 Proof of Theorem 4.1.2

Theorem 4.1.2 states two results: the turn-level teacher-student divergence drift, which concerns the Fisher-Rao geometry of the policy simplex at a single fixed state, and the policy-level teacher-student divergence drift, which aggregates the turn-level drift over the visits to that state under the frozen-occupancy assumption. We prove the two results separately, in Appendix A.4.1 and in Appendix A.4.2.

A.4.1 Proof of Eq. (10)

Setup. Fix a state s, let $a \in \mathcal{A}_s$ be the sampled response of Theorem 4.1.2 and b a generic response, assume $\pi_\theta(b \mid s) > 0$ and $\pi_T(b \mid s) > 0$ for all $b \in \mathcal{A}_s$, and write the turn-level divergence of Eq. (4) as a sum over the finite response set,

$$D_{\mathrm{turn}}(s) = \sum_{b \in \mathcal{A}_s} \pi(b \mid s)\, L(b \mid s),$$

where $L(b \mid s)$ is the divergence score of Eq. (5), the teacher-student divergence analogue of "surprise", whose expectation under $\pi(\cdot \mid s)$ is $D_{\mathrm{turn}}(s)$.

Step 1: Fisher-Rao gradient formula. Here grad_F is the Fisher-Rao gradient on the simplex $\Delta^\circ(\mathcal{A}_s)$ with the metric g_π and the tangent space $T_\pi\Delta^\circ(\mathcal{A}_s)$ of Appendix A.2. For any smooth function $f : \Delta^\circ(\mathcal{A}_s) \to \mathbb{R}$,

$$\mathrm{grad}_F f(\pi) = \pi \odot \big(\nabla_\pi f - (\pi^\top \nabla_\pi f)\, \mathbf{1}\big), \tag{27}$$

where $\odot$ is the Hadamard product, $\mathbf{1}$ the all-ones vector, and $\nabla_\pi f$ the vector of partial derivatives of f with respect to the coordinates $\pi(b \mid s)$, so that

$$\pi^\top \nabla_\pi f = \sum_{b \in \mathcal{A}_s} \pi(b \mid s)\, \frac{\partial f}{\partial \pi(b \mid s)}.$$

Indeed, for any tangent vector $\xi \in T_\pi\Delta^\circ(\mathcal{A}_s)$ with $\mathbf{1}^\top \xi = 0$,

$$\begin{aligned} g_\pi\big(\pi \odot (\nabla_\pi f - (\pi^\top \nabla_\pi f)\mathbf{1}),\, \xi\big) &= \sum_{b \in \mathcal{A}_s} \frac{\pi(b \mid s)\big(\frac{\partial f}{\partial \pi(b|s)} - \pi^\top \nabla_\pi f\big)\xi_b}{\pi(b \mid s)} \\ &= \sum_{b \in \mathcal{A}_s} \frac{\partial f}{\partial \pi(b \mid s)}\, \xi_b - (\pi^\top \nabla_\pi f) \sum_{b \in \mathcal{A}_s} \xi_b = \nabla_\pi f^\top \xi = df_\pi[\xi], \end{aligned}$$

so the displayed vector is exactly the Riemannian gradient of f under the Fisher-Rao metric, where the second line uses $\sum_{b \in \mathcal{A}_s} \xi_b = 0$; this holds for any f.

Step 2: Fisher-Rao gradients of the two functionals. We apply Eq. (27) to the two functionals that enter the drift. First, write $J_a(\pi) = A^{\mathrm{OPD}}(a, s) \log \pi(a \mid s)$, with $A^{\mathrm{OPD}}(a, s)$ constant (stop-gradient negated divergence score, Proposition 4.1.1; the pure OPD update carries no additional regularizer). Its partial derivative and its directional sum are

$$\begin{aligned} \frac{\partial J_a}{\partial \pi(b \mid s)} &= \frac{A^{\mathrm{OPD}}(a, s)\, \mathbf{1}\{b = a\}}{\pi(a \mid s)}, \qquad b \in \mathcal{A}_s, \\ \pi^\top \nabla_\pi J_a &= A^{\mathrm{OPD}}(a, s), \end{aligned}$$

so its Fisher-Rao gradient has coordinates

$$\begin{aligned} \big(\mathrm{grad}_F\, J_a(\pi)\big)_b &= \pi(b \mid s)\Big(\frac{A^{\mathrm{OPD}}(a, s)\, \mathbf{1}\{b = a\}}{\pi(a \mid s)} - A^{\mathrm{OPD}}(a, s)\Big) \\ &= A^{\mathrm{OPD}}(a, s)\big(\mathbf{1}\{b = a\} - \pi(b \mid s)\big), \qquad b \in \mathcal{A}_s. \end{aligned} \tag{28}$$

Next, for the turn-level teacher-student divergence, $D_{\mathrm{turn}}(s) = \sum_{b\in\mathcal{A}_s} \pi(b \mid s)\, L(b \mid s)$ with the teacher held fixed, the corresponding partial derivative and directional sum are

$$\frac{\partial D_{\mathrm{turn}}(s)}{\partial \pi(b \mid s)} = L(b \mid s) + 1, \qquad b \in \mathcal{A}_s,$$
$$\pi^\top \nabla_\pi D_{\mathrm{turn}}(s) = \sum_{b\in\mathcal{A}_s} \pi(b \mid s)\big(L(b \mid s) + 1\big) = D_{\mathrm{turn}}(s) + 1,$$

so

$$\begin{aligned}\big(\mathrm{grad}_F\, D_{\mathrm{turn}}(\pi)\big)_b &= \pi(b \mid s)\big(L(b \mid s) + 1 - (D_{\mathrm{turn}}(s) + 1)\big) \\ &= \pi(b \mid s)\big(L(b \mid s) - D_{\mathrm{turn}}(s)\big), \qquad b \in \mathcal{A}_s. \end{aligned} \tag{29}$$

Step 3: inner product. Combining Eqs. (28) and (29),

$$\begin{aligned} D^{\mathrm{turn}}_{\mathrm{OPD}}(a; s) &= g_\pi\Big(\big(\pi(b \mid s)(L(b \mid s) - D_{\mathrm{turn}}(s))\big)_{b\in\mathcal{A}_s}, \\ &\qquad \big(A^{\mathrm{OPD}}(a, s)(\mathbf{1}\{b = a\} - \pi(b \mid s))\big)_{b\in\mathcal{A}_s}\Big) \\ &= A^{\mathrm{OPD}}(a, s) \sum_{b\in\mathcal{A}_s} \big(L(b \mid s) - D_{\mathrm{turn}}(s)\big)\big(\mathbf{1}\{b = a\} - \pi(b \mid s)\big) \\ &= A^{\mathrm{OPD}}(a, s)\Big[L(a \mid s) - D_{\mathrm{turn}}(s) - \sum_{b\in\mathcal{A}_s} \pi(b \mid s)\big(L(b \mid s) - D_{\mathrm{turn}}(s)\big)\Big] \\ &= A^{\mathrm{OPD}}(a, s)\big(L(a \mid s) - D_{\mathrm{turn}}(s)\big), \end{aligned} \tag{30}$$

where the penultimate step uses $\sum_{b\in\mathcal{A}_s} \pi(b \mid s)\, L(b \mid s) = D_{\mathrm{turn}}(s)$ and $\sum_{b\in\mathcal{A}_s} \pi(b \mid s) = 1$. This is exactly Eq. (10), which proves the turn-level teacher-student divergence drift of Theorem 4.1.2.

A.4.2 Proof of Eq. (11)

Setup. Fix the state s and the sampled response a of Appendix A.4.1, and recall the trajectory-level decomposition of Eq. (6), which writes the policy-level divergence as an expectation over turns,

$$D_{\mathrm{policy}}(\pi) = \mathbb{E}_{s_0\sim\mathcal{D},\, \tau\sim P_\theta(\cdot|s_0)}\Big[\sum_{t=0}^{T-1} D_{\mathrm{turn}}(s_t)\Big], \tag{31}$$

where $\tau = \{(s_0, a_0), \ldots, (s_{T-1}, a_{T-1})\}$ is the rollout, $P_\theta(\cdot \mid s_0)$ its distribution, and $D_{\mathrm{turn}}(s_t)$ the turn-level divergence of Eq. (4) at the visited state s_t. Two consequences of the frozen-occupancy assumption of Theorem 4.1.2 are used throughout. *(i) Fixed rollout:* no gradient passes through the rollout distribution P_θ, so the expectation acts as a fixed linear combination of its summands. *(ii) State-s marginal:* the Fisher-Rao derivative is taken with respect to the marginal $\pi(\cdot \mid s)$ alone, holding every other marginal $\pi(\cdot \mid s_t)$ with $s_t \neq s$ fixed. For brevity, write

$$p_t := \mathbb{P}_{s_0\sim\mathcal{D},\, \tau\sim P_\theta(\cdot|s_0)}\big[s_t = s\big]$$

for the visit probability of state s at turn t, i.e. the weight that multiplies the turn-t term of Eq. (11).

Step 1: separating the visits to s. Split every summand of Eq. (31) according to the state it is evaluated at,

$$D_{\mathrm{turn}}(s_t) = \mathbf{1}\{s_t = s\}\, D_{\mathrm{turn}}(\pi) + \mathbf{1}\{s_t \neq s\}\, D_{\mathrm{turn}}(s_t),$$

where the first term is the functional $D_{\mathrm{turn}}(\pi)$ of Eq. (10), whose argument is the state-s marginal alone, and the second term depends on π only through the marginal of another state. Collecting the second group into

$$C(\pi) := \mathbb{E}_{s_0\sim\mathcal{D},\, \tau\sim P_\theta(\cdot|s_0)}\Big[\sum_{t=0}^{T-1} \mathbf{1}\{s_t \neq s\}\, D_{\mathrm{turn}}(s_t)\Big]$$

and substituting the split into Eq. (31) gives

$$D_{\text{policy}}(\pi) = \mathbb{E}_{s_0\sim\mathcal{D},\,\tau\sim P_\theta(\cdot|s_0)}\Big[\sum_{t=0}^{T-1}\Big(\mathbf{1}\{s_t = s\}\, D_{\text{turn}}(\pi) + \mathbf{1}\{s_t \neq s\}\, D_{\text{turn}}(s_t)\Big)\Big]$$
$$= \mathbb{E}_{s_0\sim\mathcal{D},\,\tau\sim P_\theta(\cdot|s_0)}\Big[\sum_{t=0}^{T-1}\mathbf{1}\{s_t = s\}\, D_{\text{turn}}(\pi)\Big] + C(\pi). \tag{32}$$

By (ii), $C(\pi)$ is constant on the simplex $\Delta^\circ(\mathcal{A}_s)$ with respect to which we differentiate, so $\operatorname{grad}_F C(\pi) = 0$. By (i) and the linearity of grad_F, the gradient of the remaining term may be taken inside the expectation and the sum, and each visit indicator then contributes its probability, $\mathbb{E}_{s_0\sim\mathcal{D},\,\tau\sim P_\theta(\cdot|s_0)}[\mathbf{1}\{s_t = s\}] = p_t$, which (i) also keeps fixed under the differentiation. Hence

$$\operatorname{grad}_F D_{\text{policy}}(\pi) = \sum_{t=0}^{T-1} p_t \operatorname{grad}_F D_{\text{turn}}(\pi), \tag{33}$$

which is the gradient-level form of Eq. (11): only the visits to s carry the derivative, each with its own visitation weight p_t that multiplies the turn-level gradient.

Step 2: inner product. By the definition of the policy-level teacher-student divergence drift in Eq. (11) and Eq. (33),

$$\begin{aligned} D_{\text{OPD}}(a; s) &:= \Big\langle \operatorname{grad}_F D_{\text{policy}}(\pi),\ \operatorname{grad}_F J_a(\pi)\Big\rangle_{\text{Fisher-Rao}} \\ &= \Big\langle \sum_{t=0}^{T-1} p_t \operatorname{grad}_F D_{\text{turn}}(\pi),\ \operatorname{grad}_F J_a(\pi)\Big\rangle_{\text{Fisher-Rao}} \\ &= \sum_{t=0}^{T-1} p_t \Big\langle \operatorname{grad}_F D_{\text{turn}}(\pi),\ \operatorname{grad}_F J_a(\pi)\Big\rangle_{\text{Fisher-Rao}} \\ &= \sum_{t=0}^{T-1} p_t\, D^{\text{turn}}_{\text{OPD}}(a; s), \end{aligned} \tag{34}$$

where the second line substitutes Eq. (33), the third uses the bilinearity of the Fisher-Rao inner product to pull out the fixed weights p_t and the finite sum—recall from (ii) that $\operatorname{grad}_F D_{\text{turn}}(\pi)$ and $\operatorname{grad}_F J_a(\pi)$ are the same two vectors for every visit t, since both depend on π only through the state-s marginal—and the last equality inserts the definition of the turn-level teacher-student divergence drift in Eq. (10).

Step 3: closed form. Finally, substituting the explicit value of $D^{\text{turn}}_{\text{OPD}}(a; s)$ proved in Appendix A.4.1,

$$D_{\text{OPD}}(a; s) = \sum_{t=0}^{T-1} \mathbb{P}_{s_0\sim\mathcal{D},\,\tau\sim P_\theta(\cdot|s_0)}\big[s_t = s\big]\, A^{\text{OPD}}(a, s)\big(L(a \mid s) - D_{\text{turn}}(s)\big), \tag{35}$$

which is exactly Eq. (11) once the weights p_t are written back in full; this completes the proof.

B Methodological Details

B.1 Methodological Analysis

B.1.1 A Baseline-Free Surrogate

The weight input that Theorem 4.1.2 identifies is the relative divergence $\rho(a \mid s)$ of Eq. (12), with L the score of Eq. (5) and D_{turn} the turn-level divergence of Eq. (4). The turn-level reverse KL against which the relative divergence is measured is not directly computable during training, since an exact evaluation of $D_{\text{turn}}(s)$ would have to enumerate the response space $\mathcal{A}_s$.

This analysis does not reconstruct that gap exactly: it converts the relative magnitude of the divergence score within the group of responses generated for the same prompt into a modulation coefficient for the response's turn.

When the old-policy log-probabilities are recomputed, two candidate inputs are available,

$$L(a \mid s) \qquad \text{and} \qquad \sum_{\ell=1}^{|a|} \delta_\ell(a, s),$$

The single-sample score and the sum of the token-level divergences $\delta_\ell(a, s)$ of Eq. (3), the latter being an estimate of the sum of the conditional means of the token-level log-ratios, built from the truncated support of Eq. (13); the distinction matters because a loss enters training through an unbiased estimator whose noise averages out over the batch, whereas a weight compares individual responses and depends on the relative divergence itself.

A single response cannot supply $D_{\mathrm{turn}}(s)$ on its own: taking the sample's own score as its reference, $\hat{D}_{\mathrm{turn}}(a \mid s) := L(a \mid s)$, makes the relative divergence of Eq. (12) vanish identically,

$$L(a \mid s) - \hat{D}_{\mathrm{turn}}(a \mid s) = 0,$$

and an independent estimate of $D_{\mathrm{turn}}(s)$ requires repeated sampling at the same state, that is, extra branches each with their own generation and teacher inference, which the online, zero-extra-inference setting rules out; the obstacle is therefore the setting and not the mathematics of the estimator. The construction analyzed below replaces the relative divergence by a surrogate free of $D_{\mathrm{turn}}(s)$, the relative magnitude of a response's divergence score within its group.

B.1.2 Licensing the Substitution

The substitution is licensed at the level of the criterion rather than of exact values: Theorem 4.1.2 writes the drift of a response as the product of its stop-gradient negated divergence score and its relative divergence, so whether training on that response decreases the divergence is decided by the sign of the product, while a positive weight rescales the product without changing its sign; the relative order of ρ across responses is therefore what the weighting follows, the ordering fixed by the criterion of Section 4.2, while the absolute scale of ρ is not needed.

Any positive map from ρ to a coefficient rescales the first-order drift without changing its sign, while the ordering by the relative divergence is fixed by the criterion of Section 4.2, and the residual freedom is the shape of the map recorded in Remark B.1.1. DT-OPD takes the map non-decreasing in the score, amplifying the turns with the largest within-group divergence scores, and the weight-reversal control of Section 5 confirms that ordering by the divergence matters.

Two boundaries keep this reading within what the first-order drift provides: the drift is first-order and local, taken at the expansion point, under frozen occupancy and along the natural-gradient direction $\mathrm{grad}_F J_a(\pi)$, so no monotone decrease of the weighted training objective is claimed; and the members of a group may come from different states, so the surrogate is not an estimator of $D_{\mathrm{turn}}(s)$, and no equivalence between the within-group relative magnitude of the score and ρ is claimed.

Remark B.1.1 Design freedom in the modulation. The first-order drift of Theorem 4.1.2 involves ρ only through its sign and through its relative magnitude, so any positive map $g : \mathbb{R} \to \mathbb{R}_{>0}$ is compatible with the first-order drift, and the criterion of Section 4.2 fixes the ordering of the weights by the relative divergence; what remains to be designed is not the absolute scale on which the score is measured, but the way in which the relative magnitude of a score within its group is converted into a coefficient for its turn, and DT-OPD chooses g non-decreasing,

$$\rho_1 \le \rho_2 \implies g(\rho_1) \le g(\rho_2),$$

so that the amplified turns are those with the largest within-group scores.

B.1.3 A Predictable Proxy for $L(a \mid s)$

To reduce the sensitivity to the particular sampled tokens, we use the predictable part of the score in place of the score itself, the sum of the token-level divergences displayed above,

$$\hat{L}(a \mid s) := \sum_{\ell=1}^{|a|} \delta_\ell(a, s), \tag{36}$$

which the Doob decomposition of Appendix A.3 separates from the sampled score as

$$L(a \mid s) = \hat{L}(a \mid s) + M_{|a|}, \qquad M_{|a|} = \sum_{\ell=1}^{|a|} \big(X_\ell(a,s) - \delta_\ell(a,s)\big),$$

with $X_\ell(a,s)$ the realized token-level log-ratios and $M_{|a|}$ the zero-mean martingale of that decomposition, the term that the proxy of Eq. (36) excludes.

Dropping the martingale keeps the mean of the score, since the discarded terms are martingale increments, and it removes the sampling noise of a particular token sequence: at each position the log-ratio splits into its conditional mean and a centered increment, and the two are uncorrelated, so the proxy retains the predictable part of the token-level variance and discards the remainder. Lemma B.1.1 states and proves both properties.

Lemma B.1.1 Properties of the Predictable Proxy. Let $\hat{L}(a \mid s)$ be the proxy of Eq. (36) and $d_\ell := X_\ell(a,s) - \delta_\ell(a,s)$ the increments of the martingale of Appendix A.3. (i) With δ_ℓ evaluated in closed form over the vocabulary, the proxy is unbiased for the turn-level divergence,

$$\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\big[\hat{L}(a \mid s)\big] = D_{\mathrm{turn}}(s),$$

and with the truncated-support estimator of Eq. (13) the identity holds up to the bias that estimator introduces; the weighting map is monotone in the estimate, so the truncation enters the weight only through it. (ii) At every token position the increment is uncorrelated with the conditional mean, and the variance of the log-ratio splits with no cross term,

$$\mathrm{Cov}\big(\delta_\ell, d_\ell\big) = 0, \qquad \mathrm{Var}\big(X_\ell\big) = \mathrm{Var}\big(\delta_\ell\big) + \mathbb{E}\big[d_\ell^2\big] \;\geq\; \mathrm{Var}\big(\delta_\ell\big).$$

Proof. (i) The tower property gives $\mathbb{E}[\delta_\ell] = \mathbb{E}[X_\ell]$ at every position, so

$$\mathbb{E}\big[\hat{L}(a \mid s)\big] = \mathbb{E}\big[L(a \mid s)\big].$$

The increments d_ℓ are martingale differences and the response length is bounded, so the stopped martingale has zero mean, and the expectation identity of Section 4.1,

$$\mathbb{E}\big[L(a \mid s)\big] = D_{\mathrm{turn}}(s),$$

transfers to the proxy.

(ii) The conditional mean δ_ℓ is measurable with respect to the autoregressive filtration $\mathcal{F}_{\ell-1}$ of Appendix A.3, hence

$$\mathbb{E}[\delta_\ell d_\ell] = \mathbb{E}\big[\delta_\ell\, \mathbb{E}[d_\ell \mid \mathcal{F}_{\ell-1}]\big] = 0, \qquad \mathbb{E}[d_\ell] = 0,$$

so the covariance vanishes; expanding $\mathrm{Var}(X_\ell) = \mathrm{Var}(\delta_\ell + d_\ell)$ with this covariance gives the exact split displayed immediately above.

B.1.4 Group-Relative Normalization

The second step sets the scale of the proxy without $D_{\mathrm{turn}}(s)$. The responses generated for the same prompt form the group G of Section 4.2, and every entry of the group is mapped to a normalized score by the construction of Eq. (15), applied after the outliers are truncated at preset group-level quantiles,

$$\tilde{D}_{i,t} = \frac{\mathrm{clip}_{q_{\mathrm{lo}},q_{\mathrm{hi}}}(\bar{D}_{i,t}) - \min_{(j,n)\in G} \mathrm{clip}_{q_{\mathrm{lo}},q_{\mathrm{hi}}}(\bar{D}_{j,n})}{\max_{(j,n)\in G} \mathrm{clip}_{q_{\mathrm{lo}},q_{\mathrm{hi}}}(\bar{D}_{j,n}) - \min_{(j,n)\in G} \mathrm{clip}_{q_{\mathrm{lo}},q_{\mathrm{hi}}}(\bar{D}_{j,n}) + \varepsilon},$$

where $\bar{D}_{j,n}$ are the length-normalized scores of Eq. (14) and $\varepsilon > 0$ is the guard against division by zero. The normalized score is read as a practical indicator of relatively low and relatively high divergence within the group.

The motivation for this step is numerical rather than statistical: a divergence score is unbounded above, and a modulator that exponentiates the raw scores directly,

$$\exp\big(+\lambda_D\, \bar{D}_{i,t}\big),$$

can overflow, whereas the normalized score of Eq. (15) lies in a bounded range. The step is a heuristic that keeps the input of the exponential in range, and no statistical role is claimed for it.

The scope of the normalized score is fixed in Section 4.2: the members of a group may come from different states, so $\tilde{D}_{i,t}$ is a relative magnitude within its group rather than an estimate of $D_{\mathrm{turn}}(s)$. Nor is any optimality claimed for the min-max form, which is a design choice with no accompanying theorem. Its invariance to a common shift is a property of the map (Lemma B.1.2(a)) rather than the reason for adopting it, and what licenses a weighting that works from the relative magnitudes of the scores alone is the criterion established above; the overall modulation scale is preserved by the self-calibrating denominator of Section 4.3, not by this step.

Lemma B.1.2 Properties of the Group-Level Normalization. Let $\bar{D}_{j,n}$ be the length-normalized scores of the group G and $\tilde{D}_{i,t}$ the normalized scores of Eq. (15). (a) A common shift of the group leaves every normalized score unchanged,

$$\bar{D}_{j,n} \longmapsto \bar{D}_{j,n} + c \text{ for all } (j,n) \in G \qquad \Longrightarrow \qquad \tilde{D}_{i,t} \text{ unchanged,}$$

(b) the composition of truncation and min-max normalization is non-decreasing,

$$\bar{D}_{j,n} \le \bar{D}_{j',n'} \qquad \Longrightarrow \qquad \tilde{D}_{j,n} \le \tilde{D}_{j',n'},$$

with equality exactly when the two clipped scores coincide, which the truncation can create at the tail, and (c) every normalized score lies in the unit interval,

$$0 \le \tilde{D}_{i,t} < 1.$$

Proof. (a) The numerator and the denominator are differences between group entries, so a common shift cancels; the quantile thresholds shift with the scores, so the clipped values they produce shift by the same constant. (b) Truncation at a fixed pair of quantiles is non-decreasing, and min-max normalization is strictly increasing on the clipped range, so the composition cannot invert the order of two distinct clipped scores and can tie them only where the truncation does. (c) After truncation every score lies between the group minimum and the group maximum, so the numerator is at most the denominator minus ε, which keeps the ratio below one.

B.1.5 Divergence-Calibrated Modulation

The final step turns the normalized score into a coefficient for the turn. Positivity is required, since a positive coefficient rescales the drift of Theorem 4.1.2 but cannot change whether it is negative or positive, and DT-OPD takes the coefficient non-decreasing in $\tilde{D}_{i,t}$, the ordering fixed by the criterion of Section 4.2, so that the turns amplified by the map are those whose divergence is relatively high within their group rather than those whose divergence is relatively low,

$$w^D_{i,t} > 0, \qquad \tilde{D}_{i,t} \le \tilde{D}_{j,n} \implies w^D_{i,t} \le w^D_{j,n}.$$

Because $\tilde{D}_{i,t}$ renormalizes a sum of non-negative token-level divergences, a larger score means larger token-level divergences, so the responses whose scores are largest within the group are those on which the student still deviates most from the teacher, and a third requirement concerns the scale of the coefficient: the normalized score carries only a within-group meaning, so a coefficient that inherited its size directly from the score would move with the distribution of the group, and the average weight of the group would drift with it. The self-calibrating denominator answers that requirement, since dividing by the group mean holds that average fixed while the map redistributes weight within the group; the overall scale of the weights is therefore set by the denominator rather than by the min-max step that produced the scores.

The construction is well posed, since the normalized score is bounded by Lemma B.1.2(c) and the guard ε keeps the denominator positive, so every coefficient is positive, bounded, and free of division by zero. No optimality is claimed for the softmax form, which is adopted to meet the requirements above rather than derived from an optimization problem, and the weight definition together with its stated properties is registered once in Section 4.3; with that, the analysis of this subsection is now complete, and no further property is claimed here.

Algorithm 1 DT-OPD

1: **Input:** student π_θ, frozen teacher π_T, prompt distribution $\mathcal{D}$, hyperparameters λ_D, ε, and the clipping quantiles $q_{\text{lo}} < q_{\text{hi}}$.
2: **Initialize:** freeze π_T; $G \leftarrow \emptyset$.
3: **for** each training step **do**
4: Sample rollouts; group responses per prompt into G.
5: Compute per-position teacher-student divergences from the distributions evaluated for OPD, over a truncated support of those distributions:

$$\delta_\ell(a_t, s_t) \leftarrow D_{\text{KL}}\big(p_\theta(\cdot \mid s_t, y_{<\ell}) \,\big\|\, p_T(\cdot \mid s_t, y_{<\ell})\big)$$

6: Length-normalize each response: $\bar{D}_{i,t} \leftarrow \frac{1}{|S_{i,t}|} \sum_{\ell \in S_{i,t}} \delta_\ell(a_t, s_t)$.
7: Truncate at the preset quantiles, $\bar{D}^{\text{cl}}_{i,t} \leftarrow \text{clip}_{q_{\text{lo}}, q_{\text{hi}}}(\bar{D}_{i,t})$, and min-max normalize within G:

$$\tilde{D}_{i,t} \leftarrow \frac{\bar{D}^{\text{cl}}_{i,t} - \min_{(j,n)\in G} \bar{D}^{\text{cl}}_{j,n}}{\max_{(j,n)\in G} \bar{D}^{\text{cl}}_{j,n} - \min_{(j,n)\in G} \bar{D}^{\text{cl}}_{j,n} + \varepsilon}$$

8: Set $w^D_{i,t} = 1$ if the clipped scores of G agree to within floating-point precision or $\lambda_D = 0$.
9: Otherwise compute the divergence-calibrated turn weights:

$$w^D_{i,t} \leftarrow \frac{\exp\big(\lambda_D \, \tilde{D}_{i,t}\big)}{\frac{1}{|G|} \sum_{(j,n)\in G} \exp\big(\lambda_D \, \tilde{D}_{j,n}\big) + \varepsilon}$$

10: Broadcast $w_t = w^D_{i,t}$ to every token of the t-th response.
11: Treat the turn weights as constants: they are computed from the current distributions and no gradient flows through them.
12: Compute the weighted DT-OPD objective:

$$\mathcal{L}_{\text{DT-OPD}}(\theta) = \mathbb{E}_{s_0 \sim \mathcal{D},\, \tau \sim P_\theta(\cdot \mid s_0)} \Big[\sum_{t=0}^{T-1} \sum_{\ell=1}^{|a_t|} w_t \cdot D_{\text{KL}}\big(p_\theta(\cdot \mid s_t, y_{<\ell}) \,\big\|\, p_T(\cdot \mid s_t, y_{<\ell})\big) \Big]$$

13: Update θ with the PPO-clipped policy-gradient step of the base OPD run on $\mathcal{L}_{\text{DT-OPD}}(\theta)$.
14: **end for**
15: **Output:** π_θ

B.2 Algorithm of Our DT-OPD

Algorithm 1 carries the exact turn-level decomposition of Proposition 4.1.1 into an operational loop: each turn's per-position teacher-student divergences are accumulated into the length-normalized score $\bar{D}_{i,t}$, so the pipeline is driven by the very divergence the OPD objective decomposes into. Group-relative truncation and min-max normalization then put all turns of a prompt group G on a common relative scale, turning the relative divergence $\rho(a \mid s)$ of Eq. (12) that Theorem 4.1.2 names as the determinant of the drift into a weight input the pipeline can act on. The self-calibrated softmax converts those scores into strictly positive turn weights $w^D_{i,t}$ that are monotone in those scores, amplifying relatively high-divergence turns and suppressing relatively low-divergence ones without changing the gradient scale, so that the weights reallocate supervision toward the group-relatively high-divergence turns while leaving, by Theorem 4.1.2, the sign of the drift induced by each response unchanged and rescaling only its magnitude.

Pipeline at a glance. At training time the weights are produced from the distributions already evaluated when the old-policy log-probabilities are recomputed. The token-level teacher-student divergences are accumulated into the length-normalized score $\bar{D}_{i,t}$ of Eq. (14), truncated at the preset group-level quantiles, min-max normalized by Eq. (15) with the guard ε, and mapped by Eq. (16) to weights whose group mean is approximately one; the weights are then broadcast to every token of the corresponding response and enter the objective of Eq. (17), which is optimized with the PPO-clipped update of Algorithm 1. The pipeline has no parameters of its own and is never differentiated through, so the per-step training budget stays at the level of the base OPD run, even though every individual turn is weighted.

C Experimental Setup Details

This section collects the details behind the experiments of Section 5: the benchmarks and their evaluation protocols, the evaluation metrics, the compared algorithms, the training hyperparameters

shared by all runs, and the prompt templates that instantiate every turn. Every student entry of Table 1 is trained and evaluated under the same frozen teacher, rollout procedure, and checkpoint-selection rule, so the differences between methods come from the training objective rather than from the evaluation setup or the teacher checkpoint.

C.1 BENCHMARKS

We evaluate on three multi-turn agentic benchmarks that cover complementary interaction patterns: embodied task execution (ALFWorld), tool-assisted knowledge retrieval (Search-QA), and web interaction (WebShop). Their typical horizons range from a few turns on Search-QA to tens of turns on ALFWorld, so the suite exercises both short tool-use loops and long-horizon decision making, and a weighting rule has to hold across all three before it can claim a general gain.

ALFWorld. A text-based embodied household simulator in which the agent completes multi-stage goals, such as finding, cleaning, heating, cooling, and placing an object, by issuing discrete actions and reading textual observations (Shridhar et al., 2020). The six evaluation categories of Table 1—Pick, Look, Clean, Heat, Cool, and Pick2—differ in the object and in the interaction pattern they require, and each episode is scored by task completion, so we report the success rate of each category. ALFWorld is the primary testbed of this work: most of its trajectories consist of many routine navigation turns and only a few decisive planning choices, which is exactly the regime in which the weighting of complete turns matters most.

Search-QA. A multi-turn retrieval-augmented question answering suite that follows the Search-R1 protocol (Jin et al., 2025) and covers both open-domain and multi-hop questions from Natural Questions (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), PopQA (Mallen et al., 2023), HotpotQA (Yang et al., 2018), 2WikiMultiHopQA (Ho et al., 2020), MuSiQue (Trivedi et al., 2022), and Bamboogle (Press et al., 2023). At every turn the agent either issues a search query or commits to a final answer, so its trajectories are short relative to ALFWorld and center on tool-assisted knowledge retrieval rather than on navigation. We report the score of every subset, in the subset order of Table 1. Search-QA has the shortest episodes of the three benchmarks, so it is the setting in which a turn weight has the fewest turns to act on and the least room to change the outcome.

WebShop. A simulated online shopping environment in which the agent must locate and purchase the product that satisfies a detailed natural-language specification, involving attributes such as size, color, and price, through search queries, clicks, and option selection (Yao et al., 2022a). It tests goal-conditioned web navigation with constraint checking over medium-length trajectories, and we report the task score and the purchase accuracy of every method in Section 5.2. All numbers in Table 1 and in Section 5.2 are percentages.

Every benchmark is scored with its native metric, defined in Appendix C.4. Within a student scale all entries share the same evaluation protocol and the same checkpoint-selection rule, so a comparison between two rows of Table 1 reflects the training objective rather than the evaluation. Because the environments differ in episode length and in the number of turns that actually carry a decision, we keep the per-category and per-subset numbers next to the averages instead of collapsing them into one single aggregate.

The reported matrix spans three student scales and the three environments, and the comparisons of Section 5 are always drawn within a single student scale and a single benchmark, since the student and the teacher differ across blocks. The same protocol covers the Qwen2.5 family comparison of Appendix D.1, which is reported block by block in the same way.

C.2 BASELINES

We compare DT-OPD with five entries that differ in their supervision source (no task-specific post-training, environment rewards, or an external teacher) and in signal granularity (trajectory-level rewards versus token- and turn-level distillation).

The entries differ along two axes. The first axis is the supervision source: Vanilla has none, GRPO uses the environment reward, OPD and TCOD use the teacher distribution, and SOD uses both, since it trains turn-level distillation weights together with a GRPO term. The second axis is the

granularity of the signal: GRPO scores a whole trajectory, whereas the distillation methods act at the token or the turn level. The turn-aware methods differ in what they attach that granularity to—the ratio between consecutive turns (SOD) or the temporal curriculum (TCOD)—which is the design space in which DT-OPD places its divergence-calibrated turn weight without environment rewards.

- **Vanilla.** The instruction-tuned student without task-specific post-training. This entry measures the agentic capability of the underlying Qwen model before any reinforcement learning or distillation, and its low success rate on ALFWorld and Search-QA quantifies how far a small model stays from the environments even after instruction tuning.
- **GRPO.** Group Relative Policy Optimization optimizes the student directly on environment rewards (Shao et al., 2024): for each prompt it samples a group of trajectories, assigns each trajectory a scalar task reward, and normalizes the rewards within the group to estimate relative advantages, which removes the need for a separate value model. In multi-turn agent tasks the reward is sparse and defined at the trajectory level, so all tokens of a trajectory receive the same coarse credit—the limitation that the remaining entries address.
- **OPD.** On-policy distillation trains the student on its own rollouts while a frozen teacher provides dense token-level supervision (Agarwal et al., 2024). Because the supervision is applied at the states the student actually visits, OPD avoids the train–test mismatch of offline imitation; every turn, however, enters the loss with the same weight regardless of how informative it is, and DT-OPD is built directly on this objective by replacing that uniform weight with $w^D_{i,t}$ of Eq. (16).
- **SOD.** Step-wise on-policy distillation is an agent-oriented variant of OPD that adapts the distillation strength at different interaction turns (Zhong et al., 2026), so that turns of different learning value receive different amounts of teacher guidance. It therefore provides more structured supervision than uniform OPD, but it reads the teacher-student divergence only through the ratio between consecutive turns, so its weights follow whether the turn-level divergence rises or falls along the trajectory, and they are optimized together with a GRPO term.
- **TCOD.** Temporal curriculum on-policy distillation schedules the temporal range of the multi-turn trajectory that is distilled (Wang et al., 2026): the student is exposed to shorter or easier temporal segments first and to longer ones later, which stabilizes multi-turn distillation but makes the outcome depend on the chosen curriculum.

The comparison separates the two ingredients of the method. Comparing GRPO with OPD shows what dense teacher supervision buys over a sparse trajectory-level reward; comparing OPD with SOD and TCOD shows what changes once the distillation signal is reweighted across a trajectory; and comparing OPD with DT-OPD isolates the divergence-calibrated turn weight at a fixed objective, because the two differ only in the weights $w^D_{i,t}$.

The Qwen3-4B and Qwen3-30B-A3B GRPO models (Yang et al., 2025a) appear as the teacher reference rows of Table 1, and the Qwen2.5-7B GRPO model (Yang et al., 2025b) as the reference row of Table 7 in Appendix D.1. All distillation entries use the same frozen teacher checkpoint where applicable, so the differences among OPD, SOD, TCOD, and DT-OPD come from how the teacher signal is weighted or scheduled rather than from the teacher itself.

C.3 Training Hyperparameters

Training is run separately for each student scale and each environment with the teacher frozen throughout. Table 5 summarizes the training setup shared by all runs, and Table 6 gives the task-specific configurations. Batch sizes, the prompt budget, and the environment step limit are chosen per benchmark; all remaining hyperparameters are shared. Rollouts are generated with vLLM (Kwon et al., 2023), and the policy is updated with the PPO-clipped policy-gradient update (Schulman et al., 2017; Shao et al., 2024).

Training protocol. Each student is paired with the teacher of its own model family: the Qwen3-0.6B and Qwen3-1.7B students (Yang et al., 2025a) are distilled from the frozen Qwen3-4B GRPO teacher, the Qwen2.5-3B student (Yang et al., 2025b) from the frozen Qwen2.5-7B GRPO teacher, and the Qwen3-4B student from the frozen Qwen3-30B-A3B GRPO teacher. Each teacher is reported as a reference row in the table listing its students, so the gap between a student and its own teacher can be read off directly rather than inferred. Runs are independent across student scales and environments, and every distillation entry of a block consumes the same rollout budget and follows

Table 5: Training setup shared by all reported methods. The teachers are frozen throughout and each is paired with the student of its own model family.

Hyperparameter	Value
Student models	Qwen3-0.6B, Qwen3-1.7B, Qwen2.5-3B, Qwen3-4B
Teacher, 0.6B/1.7B students	frozen Qwen3-4B GRPO
Teacher, 4B student	frozen Qwen3-30B-A3B GRPO
Teacher, 3B student	frozen Qwen2.5-7B GRPO
Actor learning rate	1×10^{-6}
KL loss	off
Group size (rollouts per prompt)	8
Max response length	512
Rollout engine	vLLM
GPU count	8 (single node)
Training steps	150
Rollout temperature	1.0 training, 0.4 validation
Thinking mode	off
Update rule	PPO-clipped policy-gradient update

the same evaluation rule, which keeps each comparison self-contained within a single model family and a single teacher, with no result pooled across its blocks.

Table 6: Task-specific hyperparameters for ALFWorld, Search-QA, and WebShop.

Hyperparameter	ALFWorld	Search-QA	WebShop
Train batch size	16	128	16
Validation batch size	128	512	128
PPO mini-batch size	256	256	64
Micro-batch size per GPU	32	32	8
Max prompt length	2048	4096	4096
Environment max steps	50	4	15

C.4 METRICS

We report a common set of metrics across the three benchmarks, all as percentages.

Success Rate (SR ↑). The task completion rate over the six ALFWorld categories, treated as a binary outcome bounded below by 0 and above by 100. An episode counts as successful when the environment emits its terminal success flag (Shridhar et al., 2020), and a category value is the fraction of successful episodes in that category. Because the six categories differ in the object and in the interaction pattern they require, we report the per-category values rather than a single aggregate.

Score (↑). The partial-credit task score, which separates a failure close to completion from one entirely off the mark, where a binary signal would hide the difference. On Search-QA it is the native subset score of the Search-R1 evaluation (Jin et al., 2025)—the fraction of questions whose final answer is judged correct; on WebShop the environment returns a continuous score in $[0, 100]$ (Yao et al., 2022a) from the attribute-level match between the purchased product and the specification; ALFWorld exposes only a binary signal, so Score is not reported on it.

Accuracy (Acc ↑). The binary counterpart of the WebShop score: the fraction of episodes whose purchased product matches the requested specification exactly, i.e., whose score reaches 100. Reporting Acc next to Score distinguishes methods that buy the right item from methods that buy a mostly, but not exactly, right one.

Fairness. All metrics are averaged over the full evaluation split, and every method shares the same decoding configuration, evaluation protocol, and checkpoint-selection rule, so a comparison between two rows reflects the training objective rather than the measurement.

C.5 Prompt Templates

The policy sees each environment through a text prompt that merges the task description, the interaction history, the current observation, and the admissible actions. Figures 2–4 give the exact templates used in all reported runs: the history-free variant is used at the first turn, and the variant with history from the second turn onward, with every placeholder substituted from the environment state before the prompt reaches the policy. The templates are shared by the student, the teacher, and all compared methods, so the differences between entries come from the training objective rather than from how the interaction is presented. All three templates follow the interleaved reasoning–acting format of ReAct (Yao et al., 2022b), and the rollout manager parses their tags into turns: `<think>` and `<action>` for the reasoning and the action on ALFWorld and WebShop, and `<think>` for the reasoning, `<search>` for a search query, `<information>` for the returned results, and `<answer>` for the final answer on Search-QA.

Prompt template for ALFWorld

First turn (no history). You are an expert agent operating in the ALFRED Embodied Environment. Your current observation is: {current_observation} Your admissible actions of the current situation are: [{admissible_actions}].

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

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

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

Figure 2: Prompt templates used for ALFWorld in all reported runs. The history-free variant is used at the first turn; the variant with history is used afterwards, with the placeholders filled from the environment state at every turn.

D Additional Experimental Results

D.1 Additional Performance Comparison

To verify that DT-OPD is effective beyond the Qwen3 models, we further evaluate it on the Qwen2.5 family. The Qwen2.5-3B student (Yang et al., 2025b) is distilled from a frozen Qwen2.5-7B GRPO teacher under the protocol of Section 5. Table 7 reports the ALFWorld and Search-QA results; the WebShop results of the same student are the second block of Table 2.

DT-OPD attains the highest average on both benchmarks: 80.5% on ALFWorld, 2.4 points above OPD and 3.2 points above SOD, and 46.3% on Search-QA, within 0.3 points of OPD and 2.2 points above TCOD. It is also the best or second-best entry in 12 of the 13 category and subset columns, the exception being Clean, where GRPO leads; its column maxima are Pick (97.2%), Pop (45.8%), and Hotp (43.3%), with Bamboogle shared with OPD at 67.3%. On both benchmarks the student remains below its own teacher, whose reference values are 81.2% on ALFWorld and 47.1% on Search-QA. This shows that DT-OPD is effective on a different model family.

Prompt template for Search-QA

First turn (no history). You are an expert agent tasked with answering the given question step-by-step. Your question: {task_description}

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

Subsequent turns (with history). You are an expert agent tasked with answering the given question step-by-step. Your question: {task_description}

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

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

Figure 3: Prompt templates used for Search-QA in all reported runs. The history-free variant is used at the first turn; the variant with history is used afterwards, with the placeholders filled from the environment state at every turn.

Prompt template for WebShop

First turn (no history). You are an expert autonomous agent operating in the WebShop e-commerce environment. Your task is to: {task_description}. Your current observation is: {current_observation}. Your admissible actions of the current situation are: [{available_actions}].

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

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

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

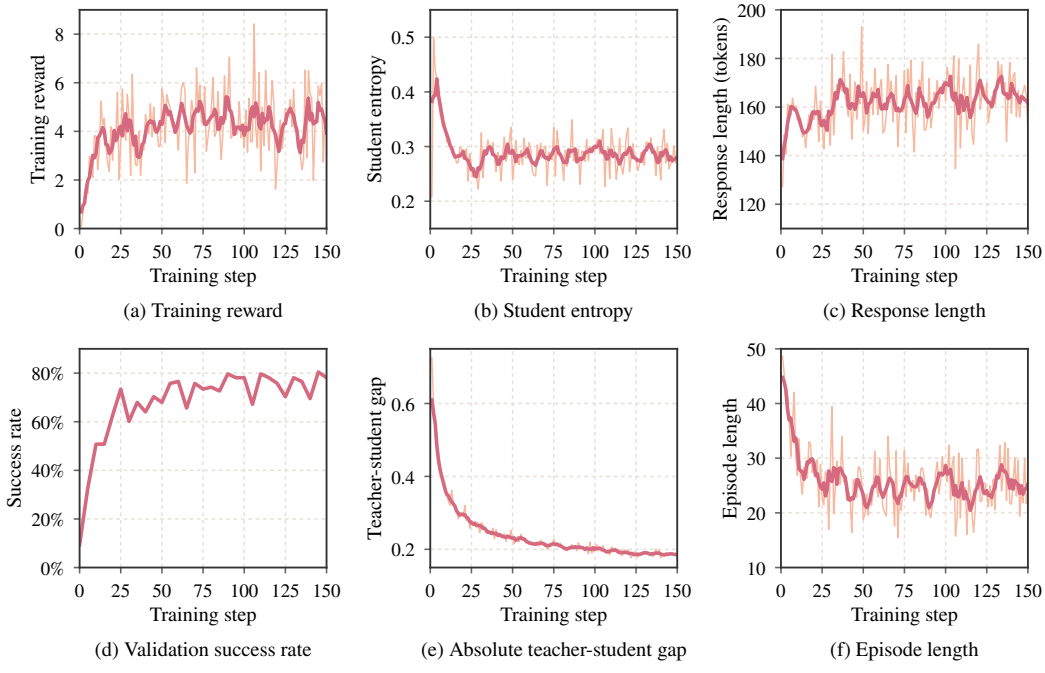
Figure 4: Prompt templates used for WebShop in all reported runs. The history-free variant is used at the first turn; the variant with history is used afterwards, with the placeholders filled from the environment state at every turn.

Table 7: Additional results on ALFWorld and Search-QA. We report success rate (SR, %), with the Qwen2.5-3B student distilled from the frozen Qwen2.5-7B GRPO teacher. Vanilla is the untrained base model. Best results are in **bold**, second best underlined.

Method	ALFWorld							Search-QA							
	Pick	Look	Clean	Heat	Cool	Pick2	Avg.	NQ	Triv	Pop	Hotp	2Wk	MuS	Bam	Avg.
Qwen2.5-7B GRPO	92.3	81.2	91.7	73.3	71.4	73.7	81.2	47.9	63.2	45.2	45.0	39.4	19.9	67.7	47.1
	Qwen2.5-3B Student (Teacher: Qwen2.5-7B GRPO)														
Vanilla	59.3	16.7	14.3	7.1	4.8	7.7	20.3	24.5	47.5	30.9	26.0	25.2	7.4	59.7	31.2
GRPO	91.2	62.5	**96.2**	61.9	65.0	47.4	75.0	39.3	60.6	41.1	37.4	36.4	15.4	26.4	36.4
OPD	96.8	81.8	84.6	46.7	64.7	**75.0**	78.1	**47.7**	**61.3**	45.7	42.9	38.3	18.7	**67.3**	46.0
SOD	96.8	**100.0**	80.8	53.3	64.7	64.3	77.3	40.7	58.6	41.5	38.8	**40.9**	17.2	33.3	43.4
TCOD	77.4	54.6	80.8	**93.3**	**70.6**	60.7	73.4	42.1	58.6	42.8	41.8	38.9	**21.5**	33.3	44.1
DT-OPD	**97.2**	87.5	81.3	64.3	66.7	72.7	**80.5**	47.4	61.0	**45.8**	**43.3**	39.4	18.9	**67.3**	**46.3**

D.2 Additional Training Dynamics

We report additional training dynamics for the Qwen3-1.7B student on ALFWorld, including the training reward, the student entropy, the response length, the validation success rate, the absolute teacher-student gap, and the episode length. The complete curves are shown in Figure 5. Over training the validation success rate improves while the absolute gap between the student and its teacher contracts from about 0.73 at the first step to about 0.18 at the final one and the episodes grow shorter, so the ALFWorld result reported for this student in Table 1 is reached along a trajectory on which the student moves toward its teacher and resolves the tasks in fewer interactions—consistent with the divergence-ordered weighting of Section 4.2, which places more supervision on the responses of larger relative divergence.



(a) Training reward (b) Student entropy (c) Response length

(d) Validation success rate (e) Absolute teacher-student gap (f) Episode length

Figure 5: Additional training diagnostics for the Qwen3-1.7B student on ALFWorld. (a) Training reward. (b) Student entropy. (c) Response length. (d) Validation success rate. (e) Absolute teacher-student gap. (f) Episode length. Thin lines are per-step values and thick lines 5-step moving averages; the validation success rate is evaluated every five steps.

D.3 Additional Ablation Study

To better understand where the gain of DT-OPD comes from, we conduct controlled ablation studies on ALFWorld with the Qwen3-1.7B student and the frozen Qwen3-4B GRPO teacher. The only difference lies in how the weights w^D of Eq. (16) are constructed or applied.

We compare the following variants:

1. **OPD:** the vanilla distillation objective without divergence-calibrated turn weighting, i.e., the token-level objective of Eq. (3),

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{s_0\sim\mathcal{D},\,\tau\sim P_\theta(\cdot|s_0)}\left[\sum_{t=0}^{T-1}\sum_{\ell=1}^{|a_t|} D_{\mathrm{KL}}\big(p_\theta(\cdot \mid s_t, y_{<\ell}) \,\big\|\, p_T(\cdot \mid s_t, y_{<\ell})\big)\right]. \tag{37}$$

2. **DT-OPD:** our full method of Section 4, which broadcasts the self-calibrated turn weights of Eq. (16) to every token of the corresponding response, $w_t := w^D_{i,t}$, so that the objective of Eq. (3) becomes

$$\mathcal{L}_{\text{DT-OPD}}(\theta) = \mathbb{E}_{s_0\sim\mathcal{D},\,\tau\sim P_\theta(\cdot|s_0)}\left[\sum_{t=0}^{T-1}\sum_{\ell=1}^{|a_t|} w^D_{i,t}\, D_{\mathrm{KL}}\big(p_\theta(\cdot \mid s_t, y_{<\ell}) \,\big\|\, p_T(\cdot \mid s_t, y_{<\ell})\big)\right], \tag{38}$$

   the objective stated as Eq. (17).

3. **DT-OPD$_{\mathrm{shuffle}}$:** computes w^D in the same way as DT-OPD but, independently for each group, randomly permutes the weights within the group before applying them to the turns,

$$\tilde{w}_{i,t} = w^D_{\sigma(i,t)}, \qquad \sigma \sim \mathrm{Uniform}(\mathfrak{S}_G), \tag{39}$$

   where σ is a uniformly random permutation of the members of the group G. This preserves the marginal distribution and the scale of w^D, but destroys the alignment between each turn and its own divergence estimate.

4. **DT-OPD$_{\mathrm{reverse}}$:** reverses the divergence-dependent weighting rule by negating the temperature λ_D in Eq. (16),

$$w^{\mathrm{rev}}_{i,t} = \frac{\exp\big(-\lambda_D\, \tilde{D}_{i,t}\big)}{\frac{1}{|G|}\sum_{(j,n)\in G}\exp\big(-\lambda_D\, \tilde{D}_{j,n}\big) + \varepsilon}, \tag{40}$$

   so that low-divergence turns are amplified and high-divergence turns are suppressed, the opposite of Eq. (16).

5. **DT-OPD$_{\mathrm{traj-norm}}$:** keeps the length normalization of Eq. (14), the softmax form of Eq. (16), and the temperature unchanged, but replaces the group-relative normalization of Eq. (15) with a trajectory-level one, rescaling each score against the turns of its own rollout instead of the group,

$$\tilde{D}^{\mathrm{traj}}_{i,t} = \frac{\bar{D}_{i,t} - \min_{n:(i,n)\in G}\bar{D}_{i,n}}{\max_{n:(i,n)\in G}\bar{D}_{i,n} - \min_{n:(i,n)\in G}\bar{D}_{i,n} + \varepsilon}, \tag{41}$$

   after which the weights follow Eq. (16) with $\tilde{D}^{\mathrm{traj}}_{i,t}$ in place of $\tilde{D}_{i,t}$.

The ablation results are reported in Table 3 of Section 5. Overall, the full DT-OPD achieves the best performance on ALFWorld, confirming that the gain comes from ordering the weights by the divergence rather than from any non-uniform weighting.

Shuffling the weights. **DT-OPD$_{\mathrm{shuffle}}$** reaches an ALFWorld average of 75.0%, below the OPD baseline (75.8%) and clearly below the full DT-OPD (78.1%). This suggests that the improvement does not come merely from introducing an additional fine-grained rescaling of turn-level losses. Instead, the key question is whether the divergence signal is assigned to the corresponding turn: once this alignment is destroyed, even if the marginal distribution and the scale of the weights are preserved, the benefit is substantially weakened.

Reversing the weighting. DT-OPD$_{\mathrm{reverse}}$ reaches an ALFWorld average of 75.0%, below the OPD baseline (75.8%) and well below the full DT-OPD (78.1%), indicating that an incorrect divergence-to-weight mapping is actively harmful. Intuitively, the reversed mapping spends the distillation budget on the turns whose responses the student already reproduces and suppresses the turns on which the teacher carries the most information—the high-divergence turns, for which $L(a \mid s) > D_{\mathrm{turn}}(s)$ and the drift of Theorem 4.1.2 is negative, so that training decreases the turn-level divergence—so the reversed map removes the divergence-decreasing pressure of exactly those turns. This result shows that what matters is not rescaling the turn-level losses, but applying the divergence-aware weights so that they follow the divergence, not reverse it.

Trajectory-level normalization. DT-OPD$_{\mathrm{traj-norm}}$ reaches an ALFWorld average of 70.3%, below the OPD baseline (75.8%) and below the full DT-OPD (78.1%). Normalizing within a single rollout draws the reference set from far fewer responses, so the relative scores inherit more sampling noise; aggregating all rollouts of the same prompt gives a stronger and more comparable reference while keeping every score anchored to the same prompt. This shows that the group-relative proxy of Eq. (15) contributes beyond the divergence ordering of the weights, and that its benefit comes from the statistics of the reference set rather than from non-uniform weighting alone.

D.4 Additional Computational Cost

This subsection compares the training-phase and validation-phase computational cost of DT-OPD with that of OPD.

Table 8: Per-step training time, validation time, and total per-step time on ALFWorld across student scales. For each student, the table reports DT-OPD and the paired same-scale OPD run whose update time is closest. The training time sums rollout generation, log-probability recomputation, teacher evaluation, weight construction, and the policy update, and excludes shared validation time.

Student	Training time (s)		Validation time (s)		Total per step (s)	
	DT-OPD	OPD	DT-OPD	OPD	DT-OPD	OPD
Qwen3-0.6B	226.1	240.5	193.0	191.2	419.1	431.7
Qwen3-1.7B	274.9	276.2	230.3	221.2	505.2	497.4
Qwen2.5-3B	307.2	301.5	248.9	257.6	556.1	559.1

Table 8 compares DT-OPD with the paired OPD run at Qwen3-0.6B, Qwen3-1.7B, and Qwen2.5-3B: the training time stays at $0.94\times$–$1.02\times$ that of OPD, and the per-step total ranges from $0.97\times$ to $1.02\times$. The only student for which DT-OPD is slower overall is Qwen3-1.7B, by 7.8 s, while the 0.6B and 3B runs are faster by 12.6 s and 3.0 s, confirming the stage-level finding of Section 5.4 that the divergence-calibrated weight adds no measurable per-step cost.