WHEN STUDENTS FAIL TO LEARN FROM TEACHERS: LEARNING DYNAMICS OF ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) has become a core technique for post-training large language models, enabling students to learn from dense token-level supervision on their own generations. However, even with a capable teacher, OPD training can be unstable, leading to excessive response length, repetitive generation, and severe performance degradation. One example is *spurious agreement*, where the student maintains high overlap with the teacher at the states visited but produces repetitive generations. Although such failures have been observed, the mechanisms behind these failures remain poorly understood. To address this gap, we investigate OPD through the lens of learning dynamics, tracing three mechanisms from local analysis to subsequent on-policy sampling. We first identify states where sampled-token OPD updates fail to correct teacher–student mismatch and characterize them using the student's mass on the teacher's top-p set. We then show that learning from the teacher is difficult at mismatched states and that their gradients can also interfere with learning at other tokens. Finally, we study how continued training on mismatched states influences on-policy sampling, suppressing the student's current modes and leading the student toward states where the teacher is unreliable. These findings help explain the benefits of supervised fine-tuning (SFT) initialization and motivate targeted correction of states where student and teacher are mismatched. We introduce state-dependent KL routing, which applies forward KL when student coverage of teacher-preferred tokens is low or a substantial fraction of teacher probability mass is underestimated. Across three Qwen3 model scales and both Base and SFT initialization, our method improves mean Avg@8 by 1.38–2.50 percentage points and mean Best@8 by 1.01–6.24 points over sampled-token OPD on mathematical reasoning benchmarks.

1 INTRODUCTION

On-policy distillation (OPD) (Agarwal et al., 2024) has become an important part of large language model (LLM) post-training, which is widely used in Qwen3 (Yang et al., 2025), Mimo (Xiao et al., 2026), DeepSeek-V4 (Xu et al., 2026a). In OPD, a student generates its own responses and learns from a teacher's token-level feedback conditioned on the resulting prefixes (Ross et al., 2011). Unlike group-based reinforcement learning (RL) algorithms, such as GRPO (Shao et al., 2024), OPD provides token-level dense reward, making it compute-efficient and effective (Lu & Lab, 2025).

However, even with a strong teacher, OPD does not always lead to effective student learning and can suffer from instability problems, such as repetitive generation and severe performance degradation (Fu et al., 2026). For example, as shown in Figure 1, OPD training leads to the phenomenon that we term *spurious agreement*: the student maintains high overlap mass with the teacher as training progresses but its generation is repetitive, leading to excessive response length.

Prior work has found these failures and proposed ways to improve the stability of OPD, such as incorporating forward KL (Fu et al., 2026) or adding a supervised fine-tuning (SFT) objective (Luo et al., 2026). In addition, other studies examine which states are useful for OPD training (Armandpour et al., 2026; Zhu et al., 2026). These include selecting informative tokens (Wang et al., 2026; Xu et al., 2026b), restricting updates within the teacher's trust region (Xing et al., 2026), and selectively applying forward KL based on conditional information such as teacher entropy (Jin et al., 2026) or advantage (Jia et al., 2026). While these studies improve the stability of OPD, they do not provide a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

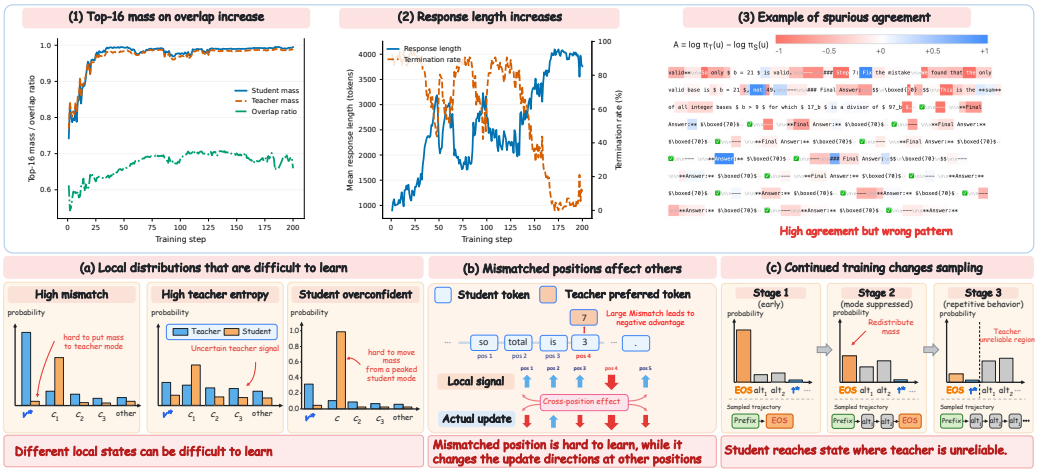


Figure 1: Learning dynamics of OPD. Overlap definitions are given in Appendix D.1.

principled explanation of how token-level teacher feedback contributes to the policy updates and how these updates affect subsequent learning.

Research Question, Proposed Work and Contributions. Based on the above discussion, this work focuses on addressing the following key research question:

> Given a strong teacher, why does a student sometimes fail to learn from the teacher's feedback during OPD, and how can we improve learning?

To answer this question, we study OPD through the lens of learning dynamics (Ren & Sutherland, 2025), from a softmax policy to practical LLM training. We first examine how sampled-token updates change the student's token distribution at the given state and identify states where teacher–student mismatch remains difficult to correct. We then study how gradients from different token positions contribute to the final distribution change, focusing on how different positions affect learning at other positions. Finally, we examine how continued training changes subsequent on-policy sampling, connecting local policy updates to the generation failures illustrated in Figure 1. Our main contributions are summarized as follows:

(1) To begin with, we study how sampled-token updates change the student's distribution at a given state. Our theoretical analysis under a single-state softmax policy characterizes the direction and magnitude of token probability changes, showing why a student may fail to learn from the teacher or even worsen after updating.

(2) We investigate how these learning dynamics affect practical LLM training. We identify states where sampled-token updates provide insufficient supervision and show how the student's probability mass on the teacher's top-p set helps characterize them. We further show that states with high mismatch can be difficult for the student to learn from while their gradients will affect learning at other positions. Finally, we examine how teacher–student mismatch affects subsequent on-policy sampling. Using EOS-token mismatch as a case, we show that continued training leads to spurious agreement.

(3) Based on the above findings, we show that SFT provides a favorable initialization for OPD and introduce state-dependent KL routing to address low student coverage of teacher-preferred tokens or underestimation of a substantial fraction of teacher probability mass. Across Qwen3 model scales and mathematical reasoning benchmarks, our approach improves training stability and performance over sampled-token OPD and the baselines.

2 Preliminaries

On-policy Distillation. The objective of on-policy distillation (OPD) is to let a student model imitate the teacher model's behavior using on-policy data. Specifically, given a prompt $x \sim \mathcal{D}_x$,

the trajectory $y = (y_1, y_2, ..., y_T)$ is sampled from the student model with a total of T generated tokens. The objective of OPD is to minimize the sequence-level reverse KL over student-generated trajectories. By the autoregressive factorization, this objective decomposes as

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{x\sim\mathcal{D}_x, y\sim\pi_\theta(\cdot|x)}\left[\sum_{t=1}^{T} D_{\mathrm{KL}}\left(\pi_\theta(\cdot \mid s_t)|\pi_T(\cdot \mid s_t)\right)\right], \tag{1}$$

where $s_t = (x, y_{<t})$ is a student-visited state.

In practice, implementations vary in how the reverse KL is computed. The methods can be categorized into three classes: (1) full-vocabulary OPD computes the divergence over the full vocabulary; (2) partial OPD restricts the sum to a subset, such as the teacher's or student's top-K tokens; and (3) sampled-token OPD uses student-sampled tokens to obtain an unbiased Monte Carlo estimator. The latter avoids an explicit vocabulary-wide KL computation and is the focus of this work.

Following Lu & Lab (2025), the resulting sampled-token OPD gradient is

$$\nabla_\theta \mathcal{L}_{\mathrm{opd}}(\theta) = -\mathbb{E}_{y\sim\pi_\theta}\left[\sum_{t=1}^{T} A_t \, \nabla_\theta \log \pi_\theta(y_t \mid s_t)\right], \tag{2}$$

where $A_t = \log \pi_T(y_t \mid s_t) - \log \pi_\theta(y_t \mid s_t)$ denotes the token-level advantage.

3 Policy Dynamics Analysis

In this section, we first present a theoretical and empirical analysis of how OPD updates change token probabilities under a single-state softmax policy.

At a fixed state s, the student policy with the softmax parametrization can be written as

$$\pi_\theta^k(a \mid s) = \frac{\exp(\theta_{s,a}^k)}{\sum_{b\in\mathcal{V}} \exp(\theta_{s,b}^k)}, \tag{3}$$

where $\mathcal{V}$ is the vocabulary set, and k is the iteration index. Let $A(s,u) := \log \pi_T(u|s) - \log \pi_\theta^k(u|s)$ denote the advantage, a single-sample SGD step on the sampled-token OPD loss updates the logits as

$$\theta_s^{k+1} \;=\; \theta_s^k \;+\; \eta\, A(s,u)\left(\mathbf{e}_u - \pi_\theta^k(\cdot|s)\right), \qquad u \sim \pi_\theta^k(\cdot|s), \tag{4}$$

where $\mathbf{e}_u \in \mathbb{R}^{|\mathcal{V}|}$ is the one-hot vector for the sampled token u.

We characterize the resulting one-step dynamics through the differences:

$$\Delta \log \pi(c \mid s) := \log \pi_\theta^{k+1}(c \mid s) - \log \pi_\theta^k(c \mid s), \qquad \Delta\pi(c \mid s) := \pi_\theta^{k+1}(c \mid s) - \pi_\theta^k(c \mid s). \tag{5}$$

Proposition 1 (Per-token training dynamics under OPD)**.** *After one update equation 4, for every $c \in \mathcal{V}$,*

$$\Delta \log \pi(c|s) = \eta\, A(s,u)\, B(c,u;s) + O(\eta^2), \quad \Delta\pi(c|s) = \pi_\theta^k(c|s)\, \Delta \log \pi(c|s) + O(\eta^2), \tag{6}$$

where

$$B(c,u;s) \;:=\; \sum_{b\in\mathcal{V}} \big(\mathbf{1}\{c=b\} - \pi_\theta^k(b|s)\big)\big(\mathbf{1}\{u=b\} - \pi_\theta^k(b|s)\big). \tag{7}$$

The proof is presented in Appendix B.1. The factorization decomposes the policy dynamics into three parts: the advantage $A(s,u)$, the probability weight $\pi_\theta^k(c|s)$, and the coupling term $B(c,u;s)$ that captures the local distribution. We next analyze Prop. 1 in terms of update direction and magnitude.

Direction of token probability changes. Let $p := \pi_\theta^k(u \mid s)$ and $S_{-u}(s) := \sum_{b\neq u} \pi_\theta^k(b \mid s)^2$. For the sampled token $c = u$,

$$B(u,u;s) = 1 - 2p + \|\pi_\theta^k(\cdot \mid s)\|_2^2 = (1-p)^2 + S_{-u}(s) \geq 0. \tag{8}$$

Therefore, with a sufficiently small learning rate η, the sampled token always moves in the direction of advantage, i.e., $\mathrm{sign}\big(\Delta\pi(u \mid s)\big) = \mathrm{sign}\big(A(s,u)\big)$. The update also redistributes probability mass among non-sampled tokens through the softmax normalization. For the token $c \neq u$,

$$B(c,u;s) = \tau(u;s) - \pi_\theta^k(c \mid s), \tag{9}$$

where $\tau(u;s) := \|\pi_\theta^k(\cdot \mid s)\|_2^2 - \pi_\theta^k(u \mid s)$. Hence, non-sampled tokens on opposite sides of $\tau(u;s)$ move in opposite directions. Unlike the token u, their movement is therefore determined not only by the advantage but also by their position in the current student distribution.

Magnitude of token probability changes. We next analyze the magnitude of token probability change. From equation 8, we have

$$\Delta\pi(u \mid s) = \eta A(s,u)\, p\big[(1-p)^2 + S_{-u}(s)\big] + O(\eta^2). \tag{10}$$

Since $0 \le S_{-u}(s) \le (1-p)^2$, the first-order update magnitude satisfies

$$|A(s,u)|p(1-p)^2 \;\le\; \lim_{\eta\to 0} |\Delta\pi(u \mid s)|/\eta \;\le\; 2|A(s,u)|p(1-p)^2. \tag{11}$$

The sampled-token probability change is therefore jointly controlled by the advantage and the token's probability. In particular, even when $|A(s,u)|$ is large, the effective probability change becomes small when the sampled token lies near either boundary: the factor p limits the magnitude of probability changes for low-probability tokens, while $(1-p)^2$ limits them for high-probability tokens.

For a non-sampled token $c \neq u$, the change of probability can be written as

$$|\Delta\pi(c \mid s)| = \eta |A(s,u)|\, \pi_\theta^k(c \mid s)\, \big|\tau(u;s) - \pi_\theta^k(c \mid s)\big| + O(\eta^2), \tag{12}$$

which depends on the advantage, token's probability and $\tau(u;s)$. In particular, low-probability tokens receive little probability change. We defer a more detailed analysis to Appendix C.

Takeaway. These analyses explain why a student may fail to learn from a strong teacher. First, tokens preferred by the teacher may change little when their student probabilities are low. Second, non-sampled tokens may change in directions that do not match the teacher's preferences. Third, the sampled-token advantage compares the teacher's and student's probabilities for only one token. It can be small even when their full distributions differ greatly.

4 Diagnosing OPD from Learning Dynamics

Building on the above analysis, we investigate why a student may fail to learn from a strong teacher in practical LLM training. First, we study local distribution dynamics to understand when sampled-token OPD updates make it hard for the student to learn from the teacher. We then examine cross-position effects, focusing on how gradients from different token positions affect actual update directions. Finally, we study on-policy sampling dynamics to understand how continued training changes the student's generation behavior. Experimental settings and details are provided in Appendix D.1.

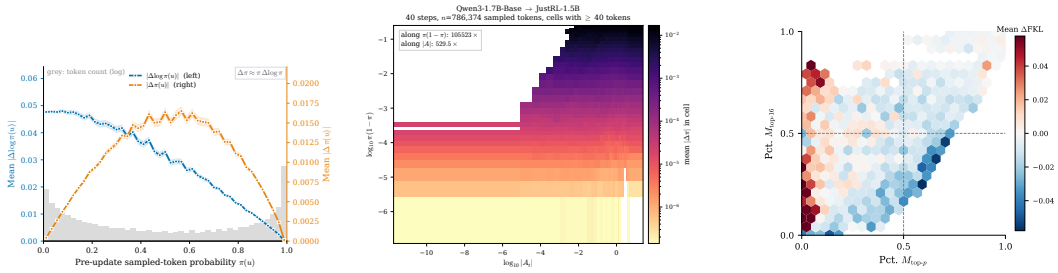
4.1 Local Distributional Dynamics

Token probability changes. The above analysis shows that the token's current probability affects the magnitude of its probability change. We first examine whether this pattern also appears in practical LLM training. As shown in Figure 2(a), low-probability sampled tokens can have large log-probability changes but small probability changes, whereas $|\Delta\pi|$ peaks at intermediate student probabilities. Figure 2(b) further shows that probability change varies with $\pi(1-\pi)$ even within similar ranges of $|A|$. Appendix D.2 extends this analysis to non-sampled tokens and compares probability- and advantage-based predictors. These results indicate that token probability changes are small when the student's current probability is either low or high. In particular, when the student assigns little probability mass to teacher-preferred tokens, sampled-token updates fail to learn the teacher's preferred token.

Top-p coverage characterizes distributional mismatch. We next examine whether these cumulative changes bring the student distribution closer to the teacher's. We measure teacher–student mismatch using forward KL (FKL). For a prefix $s_i = (x, y_{<i})$, we define

$$\mathrm{FKL}_i = \sum_{v\in\mathcal{V}} \pi_T(v \mid s_i)(\log \pi_T(v \mid s_i) - \log \pi_\theta(v \mid s_i)). \tag{13}$$

To isolate changes in the student's distribution from changes in the states it visits, we use a held-out set of fixed teacher-generated prefixes. We evaluate the student on the same prefixes before training and after multiple OPD updates, and report the change $\Delta\mathrm{FKL}_i = \mathrm{FKL}_i^{\mathrm{after}} - \mathrm{FKL}_i^{\mathrm{initial}}$. Negative values indicate improved alignment, while positive values indicate deterioration.



(a) (b) (c)

Figure 2: **Local distributional dynamics under OPD.** (a) Log-probability and probability-mass changes after one training step grouped by student probability. (b) Mean $|\Delta\pi|$ within joint bins of $|A|$ and $\pi(1-\pi)$. (c) Mean ΔFKL jointly grouped by student coverage of the teacher's top-p ($p = 0.9$) and top-K (K=16) sets. Blue indicates improved alignment; red indicates deterioration.

To better understand teacher–student mismatch, we measure how much student probability lies on the teacher's high-probability tokens. Let $\mathcal{S}_p(s)$ be the smallest set of highest-probability teacher tokens whose cumulative teacher mass reaches p, and let $\mathcal{S}_K(s)$ contain the teacher's K highest-probability tokens. We define the student's coverage of these sets as

$$M_{\text{top-}p} = \sum_{c \in \mathcal{S}_p(s)} \pi_\theta(c \mid s), \quad M_{\text{top-}K} = \sum_{c \in \mathcal{S}_K(s)} \pi_\theta(c \mid s). \tag{14}$$

Low coverage indicates that the student assigns little probability mass to teacher-preferred tokens. Unlike top-K, top-p adjusts the number of selected tokens based on the teacher's distribution.

Figure 2(c) shows that deterioration concentrates at low top-p coverage across a broad range of top-K coverage. This suggests that top-p coverage is more informative for capturing the mismatch between student and teacher. We therefore use top-p coverage in the following analyses.

Characterizing states that are difficult to learn from. We next study which states are difficult for the student to learn from. We measure the change in FKL at fixed held-out states before and after OPD and analyze states in the highest and lowest initial FKL quartiles separately. We characterize these states using student and teacher entropy, and the student's coverage of the teacher's top-p set.

We first compare two students with the same teacher, `JustRL-1.5B`, an RL-trained version of `DeepSeek-R1-Distill-Qwen-1.5B` (R1-Distill-1.5B). As shown in Figure 3(e,j), the R1-Distill-1.5B student maintains high coverage on both training and held-out states, while the Qwen3-1.7B-Base student has substantially lower coverage. As a result, for the Qwen student, FKL increases at states with high mismatch and low student entropy (Figure 3(a)). At states with low top-p coverage, even when FKL decreases, the improvement is small relative to the large initial mismatch distance (Figure 3(b)). When initial mismatch is low, changes are generally smaller, and the student becomes worse in regions with high teacher entropy. In contrast, the R1-Distill-1.5B student learns well at different states.

As a complementary comparison, we fix the student and vary the teacher (Fig. 4). When the student assigns more probability mass to the teacher's top-p set, this coverage further increases during training and AMC23 accuracy improves substantially. With lower coverage, the student shows little improvement. This further supports the importance of student coverage for learning from teacher feedback.

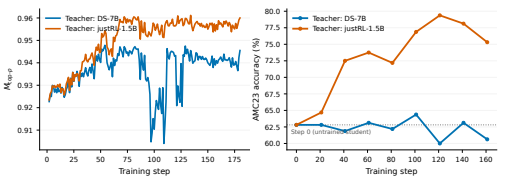


Figure 4: Same student (R1-Distill-1.5B) with different teachers, DeepSeek-Distill-Qwen-7B (DS-7B) and JustRL-1.5B.

These results demonstrate the importance of student coverage of the teacher's distribution for learning under sampled-token OPD. When student coverage is high, it generally learns well from the teacher's supervision. When student and teacher have substantial mismatch, the gap can increase at states with high teacher entropy and low student entropy. At low-coverage states, FKL may

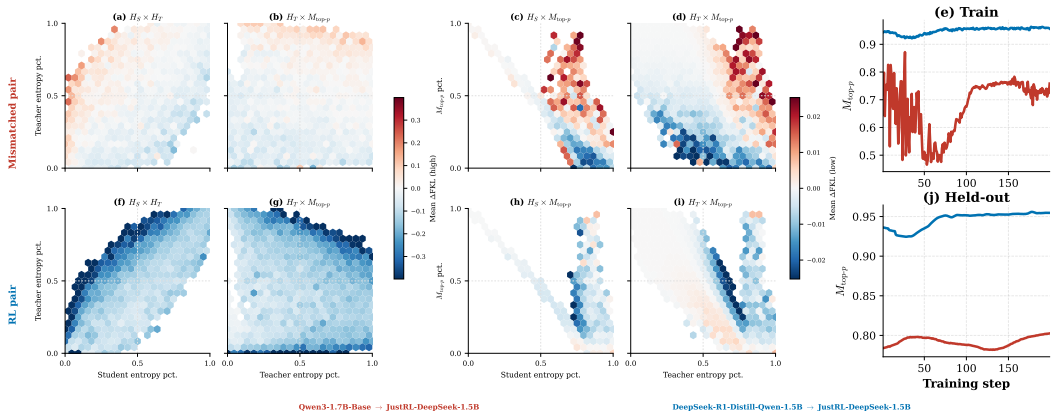


Figure 3: **Local learning dynamics under two teacher–student pairs.** Rows show the mismatched and RL-aligned pairs. Columns 1–2 and 3–4 show changes in FKL at fixed held-out states with high and low initial FKL, respectively. The final column tracks $M_{\text{top-}p}$ during training.

decrease, but the improvement is small relative to the large initial mismatch. We next examine how cross-position contributions can further affect the updates induced by teacher feedback.

4.2 Cross-Position Effects

The softmax analysis treats logits at each state as independent parameters; however, this assumption no longer holds in an LLM. In an LLM, logits across positions are produced by shared parameters, so an update at one position also affects others. We study how these effects change the student's actual updates during OPD.

Token update direction. In Prop. 1, for a sufficiently small update, the sampled token changes in the direction of its advantage, e.g., $\text{sign}(\Delta \log \pi_S(u_i)) = \text{sign}(A_i)$. However, in LLM training, the two signs agree for only around 50% across different teacher–student pairs (Figure 5(a)). To understand this difference, we vary the number of token positions included in an update. The sign agreement between advantage and change is close to 100% when only one position is included, but decreases as more positions are added (Figure 5(b)). This indicates that the gradients from other positions can change the update direction (see Appendix D.3 for additional analysis).

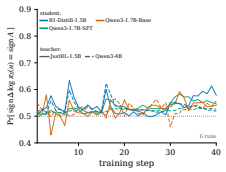
For a batch of tokens $\mathcal{B}$, let $h_i = \nabla_\theta \log \pi_\theta(u_i \mid s_i)$ denote the log-probability gradient. A single SGD update gives $\Delta\theta = \eta \sum_{j \in \mathcal{B}} A_j h_j$. We define the first-order effect of position j on position i as $M_{ij} = \eta A_j \langle h_i, h_j \rangle$. The resulting change at position i is

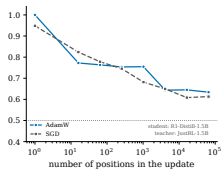
$$\Delta \log \pi_\theta(u_i \mid s_i) = \eta A_i \|h_i\|^2 + \sum_{j \neq i} M_{ij} + O(\eta^2). \tag{15}$$

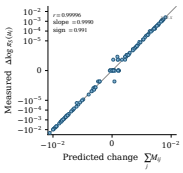
The first term follows the direction of A_i, while the total cross-position effect can reinforce, cancel, or reverse this direction. To test the first order approximation in equation 15, we compare the predicted change $\sum_j M_{ij}$ with the measured change after one SGD step in Figure 5(c).

Effects of mismatched positions. We next examine which positions have a large effect on updates at other positions. Interaction matrices across model pairs are provided in Appendix D.3. Specifically, Figure 5(d) shows the five positions with the largest cross-position influence in a selected batch under `Qwen3-1.7B-Base` and `JustRL-1.5B`. Among these positions, the termination token has a large negative advantage and the largest influence on others.

Takeaway. These results show that teacher–student mismatch at one position can affect learning at other positions through shared model parameters. Specifically, at a position where the student and teacher strongly disagree, the student cannot learn from the teacher even with a large negative advantage. Moreover, the gradient at this state will affect learning at other positions. We next examine how these realized updates affect subsequent sampling in training.







Rank	Token	A	Weight
1	`<|endoftext|>`	−22.53	31.81%
2	`Rightarrow`	3.76	10.09%
3	`that`	2.09	6.20%
4	`become`	−2.61	3.55%
5	`terms`	−1.35	3.03%

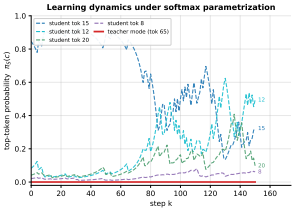
Figure 5: **Cross-position effects on update directions.** (a) Sign agreement between A_i and $\Delta \log \pi_\theta(u_i \mid s_i)$ across training steps. (b) Sign agreement as the number of positions included in an update increases. (c) First-order predictions $\sum_j M_{ij}$ versus measured changes after one SGD step. (d) Top 5 tokens by absolute cross-position influence $\sum_{i \neq j} |M_{ij}|$. Weights are these influences normalized to sum to 100% over the selected 200 positions.

4.3 Sampling and Trajectory Dynamics

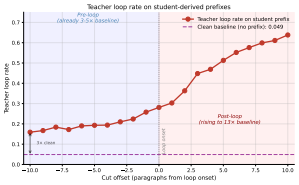
The previous sections show that under sampled-token OPD training, the student may fail to learn from the teacher at some states or even move further away from the teacher distribution. Positions with substantial teacher–student mismatch are difficult to learn from, while their gradients can still affect updates at other positions. Under on-policy sampling, these updates can also change which states the student visits next. We now examine how continued training affects subsequent sampling and generation behavior.

Student mode suppression before learning the teacher's preferred token. We first examine what happens when the student fails to learn a teacher-preferred token. We consider two categorical policies, where they have substantial mismatch. We repeatedly apply sampled-token OPD updates and track the student's top-4 tokens and the teacher's preferred token in Figure 6a. As training progresses, the student's mode gradually disappears as probability mass shifts to other tokens, while the probability of the teacher's preferred token remains near zero. This shows that sampled-token OPD suppresses the student's existing mode without learning the teacher's preferred token.

EOS-token mismatch and generation behavior. We next study termination token mismatch as a concrete example of how sampled-token OPD at this state affects subsequent sampling. The student (`Qwen3-1.7B-Base`) uses `<|endoftext|>` to end a response, while the teacher (`Qwen3-8B`) uses `<|im_end|>`. During training, the student's termination token is suppressed, while the teacher's termination token remains difficult to learn. As a result, response length increases and completion rate decreases within the generation budget, while the probability mass on the overlap between the student and teacher remains high (Figure 1). This illustrates *spurious agreement*: the student enters longer, repetitive trajectories while maintaining high local overlap with the teacher at the visited prefixes.



(a) Categorical policy.



(b) Teacher continuation.

Figure 6: **On-policy sampling and trajectory dynamics. (a)** Token probabilities under a categorical policy. **(b)** Teacher loop rate when continuing student prefixes around the first repeated paragraph. Loop rate is the fraction of sequences in which at least five consecutive tokens are repeated at least twice.

To understand why high local overlap persists between the student and teacher despite repetitive generation, we ask the teacher to continue student-generated prefixes before and after repetition begins. Figure 6b shows that the teacher rarely repeats content when generating from the original prompt. Its loop rate increases when it continues student-generated prefixes near and after the start of repetition. Thus, continued training can lead the student to visit states where the teacher is also more likely to repeat. At these states, the teacher cannot provide useful feedback for correcting the student's generation.

5 IMPLICATIONS FOR OPD TRAINING

The analysis points to two ways to improve learning from teacher feedback: reduce teacher–student mismatch before OPD, and provide distribution-level feedback at states where sampled-token updates give weak correction. These motivate SFT initialization and state-dependent KL routing, respectively.

5.1 UNDERSTANDING SFT INITIALIZATION FOR OPD

Prior work (Li et al., 2026) has shown that SFT on high-quality data or teacher trajectories can improve subsequent OPD. Our analysis helps explain why SFT initialization helps at three levels. (1) It increases the probability of teacher-preferred tokens, making them more likely to be sampled and updated. (2) It reduces the cross-position influence from hard-to-learn tokens, limiting their effects on other positions' update directions. (3) It reduces repeated updates at mismatched states, helping preserve existing student modes and avoid states where the teacher is unreliable.

5.2 STATE-DEPENDENT KL ROUTING

Although SFT can reduce teacher–student mismatch, learning from the teacher may remain difficult at some states. We use coverage in equation 14 to identify insufficient student mass on teacher-preferred tokens. To additionally measure the fraction of teacher top-p mass assigned to tokens underestimated by the student, we define

$$V(\theta; s_t) = \sum_{c \in \mathcal{S}_p(s_t)} \hat{\pi}_T(c \mid s_t) \mathbf{1}[\pi_T(c \mid s_t)/\pi_\theta(c \mid s_t) > \varepsilon], \tag{16}$$

where $s_t = (x, y_{<t})$ and $\hat{\pi}_T$ is the teacher distribution normalized over its top-p set. The threshold $\varepsilon > 1$ identifies tokens whose teacher probability exceeds ε times their student probability. Thus, $V(\theta; s_t)$ measures the fraction of teacher mass within this set assigned to such tokens.

We apply approximate top-K FKL withnout normalization when the student has low overall coverage or underestimates a substantial fraction of the teacher's probability mass:

$$\begin{aligned} r(s_t) &= \mathbf{1}[V(\theta; s_t) > \varepsilon_2 \ \vee \ M_{\text{top-}p}(s_t) < \varepsilon_3], \\ \mathcal{L}_{\text{route}}(\theta; s_t) &= (1 - r(s_t))\mathcal{L}_{\text{OPD}}(\theta; s_t) + r(s_t) \sum_{c \in \mathcal{S}_K(s_t)} \pi_T(c \mid s_t) \log \frac{\pi_T(c \mid s_t)}{\pi_\theta(c \mid s_t)}. \end{aligned} \tag{17}$$

Here, $\varepsilon_2 \in [0, 1]$ is the threshold on underestimated teacher mass, and $\varepsilon_3 \in [0, 1]$ is the minimum student coverage threshold. $\mathcal{L}_{\text{OPD}}(\theta; s_t)$ denotes the per-state reverse-KL objective, and $\mathcal{S}_K(s_t)$ is the teacher's top-K set used for efficient FKL computation. Teacher probabilities are not renormalized over the top-K set. The method retains sampled-token OPD by default.

6 EXPERIMENT

Models and Evaluation. We conduct experiments using three Qwen3 (Yang et al., 2025) models: Qwen3-1.7B-Base, Qwen3-4B-Base, and Qwen3-8B-Base. We use Qwen3-8B as the teacher model in both settings: non-thinking mode for OPD directly from Base initialization, and thinking mode for OPD from SFT initialization. We use the DAPO dataset (Yu et al., 2026) for OPD training across all student sizes. For SFT, we extract 400K prompt–response pairs from `OpenThoughts3-1.2M`, comprising 50K distinct prompts with eight responses each. We evaluate on mathematical reasoning benchmarks: AIME24-26 (MMA), and AMC23 (MMA, 2023). We report Avg@8, the mean accuracy across the eight responses, and Best@8, the bootstrap estimate of best-of-eight accuracy obtained by resampling with replacement. For more details, we defer to Appendix E.

Baselines. We compare our method against three baselines: (i) Sampled-token OPD (Lu & Lab, 2025); (ii) EOPD (Jin et al., 2026), which uses forward KL when the teacher has high entropy; (iii) AOPD (Jia et al., 2026), which uses FKL when the sampled-token advantage is non-positive.

Benchmark	Metric	Qwen3-1.7B-Base					Qwen3-4B-Base					Qwen3-8B-Base					Teacher Qwen3-8B
		SFT	OPD	EOPD	AOPD	Ours	SFT	OPD	EOPD	AOPD	Ours	SFT	OPD	EOPD	AOPD	Ours	
(a) Qwen3-8B non-thinking teacher																	
AMC23	Avg@8	–	38.44	38.75	33.12	**45.00**	–	53.75	54.37	55.94	**58.75**	–	60.94	54.38	60.31	**61.56**	69.38
	Best@8	–	64.22	**70.68**	58.09	70.65	–	79.12	79.67	**83.89**	80.74	–	79.90	**84.87**	**84.87**	83.89	89.58
AIME24	Avg@8	–	8.75	9.58	6.25	**12.08**	–	15.83	17.08	15.83	**18.75**	–	20.42	21.67	**22.50**	**22.50**	25.83
	Best@8	–	20.68	18.70	19.45	**24.60**	–	26.25	26.67	**34.10**	32.94	–	32.08	37.28	38.01	**41.77**	42.84
AIME25	Avg@8	–	**9.58**	7.08	5.42	**9.58**	–	15.00	16.67	16.25	**18.75**	–	17.50	**18.33**	16.25	17.50	20.42
	Best@8	–	14.31	13.25	17.16	**21.89**	–	27.70	28.78	25.91	**32.82**	–	27.89	29.58	20.99	**31.60**	36.71
AIME26	Avg@8	–	**6.25**	5.42	4.58	5.83	–	**15.42**	13.33	10.83	13.75	–	14.58	13.75	12.50	**18.75**	15.83
	Best@8	–	**19.66**	14.80	12.55	12.12	–	24.44	24.50	25.29	**28.05**	–	26.31	26.24	24.72	**33.91**	33.44
Mean	Avg@8	–	15.76	15.21	12.34	**18.12**	–	25.00	25.36	24.71	**27.50**	–	28.36	27.03	27.89	**30.08**	32.87
	Best@8	–	29.72	29.36	26.81	**32.32**	–	39.38	39.91	42.30	**43.64**	–	41.55	44.49	42.15	**47.79**	50.64
(b) Qwen3-8B thinking teacher																	
AMC23	Avg@8	**46.56**	44.69	42.50	45.31	45.94	74.69	77.81	77.50	77.81	**80.00**	82.19	83.12	81.88	82.19	**85.31**	91.56
	Best@8	**72.50**	66.70	61.68	67.38	64.92	92.50	90.41	91.30	90.53	**93.16**	**95.00**	93.14	90.79	91.60	94.06	99.77
AIME24	Avg@8	10.83	13.33	11.67	12.50	**14.58**	42.08	47.50	44.58	47.50	**48.75**	46.25	57.92	**60.83**	58.33	59.58	67.50
	Best@8	30.00	26.95	22.95	26.17	**34.45**	**73.33**	66.27	67.64	71.16	70.41	73.33	**79.84**	76.22	75.10	79.77	78.23
AIME25	Avg@8	13.75	16.67	13.75	**18.75**	17.08	34.17	37.08	**42.08**	36.25	37.50	37.08	40.42	40.83	40.83	**41.67**	53.33
	Best@8	23.33	28.44	27.80	**29.81**	28.34	**60.00**	57.32	58.77	56.92	57.25	56.67	61.30	60.23	59.05	**62.31**	73.68
AIME26	Avg@8	9.17	8.75	11.25	**15.42**	14.17	37.08	40.83	42.08	41.25	**42.50**	46.25	50.83	47.08	48.33	**52.08**	55.83
	Best@8	26.67	18.43	28.01	**29.60**	26.12	**60.00**	54.76	52.94	55.57	57.15	66.67	69.39	62.56	**72.07**	71.56	71.37
Mean	Avg@8	20.08	20.86	19.79	**23.00**	22.94	47.01	50.81	51.56	50.70	**52.19**	52.94	58.07	57.66	57.42	**59.66**	67.06
	Best@8	38.13	35.13	35.11	38.24	**38.46**	**71.46**	67.19	67.66	68.55	69.49	72.92	75.92	72.45	74.46	**76.93**	80.76

Table 1: Math reasoning results with Qwen3-8B as teacher. (a) Base initialization, non-thinking teacher. (b) Shared SFT initialization per model size, thinking teacher. Our routing settings are $p = 0.9$, $K = 16$, $\varepsilon_1 = 6$, $\varepsilon_2 = 0.3$, and $\varepsilon_3 = 0.2$. Bold indicates the best result.

Main Result. Table 1 compares methods within each initialization setting, with different teacher modes across panels. From Base initialization, our method achieves the highest mean Avg@8 and Best@8 at all three scales. Relative to sampled-token OPD, mean Avg@8 improves by 2.36/2.50/1.72 percentage points and mean Best@8 by 2.60/4.26/6.24 points for 1.7B/4B/8B, respectively.

With SFT initialization, the corresponding gains are 2.08/1.38/1.59 points in mean Avg@8 and 3.33/2.30/1.01 points in mean Best@8. Our method achieves the highest mean Best@8 among OPD methods at all three scales, and the highest mean Avg@8 at 4B and 8B, with a comparable result at 1.7B. For the 4B student, all OPD variants improve Avg@8 but reduce Best@8 relative to the SFT checkpoint.

Figure 7 shows the training dynamics curve of different methods, includes All-FKL, which applies top-K FKL at every position. OPD and EOPD exhibit *spurious agreement*: top-16 overlap and teacher-support coverage remain high despite collapsing completion rates. Our method maintains stable completion and the highest coverage of the teacher's top-p support, achieving the best performance while using FKL at the fewest positions among the adaptive methods. We also compare against All-FKL, which applies FKL at every position but does not achieve better performance and incurs higher memory costs as discussed in Jin et al. (2026).

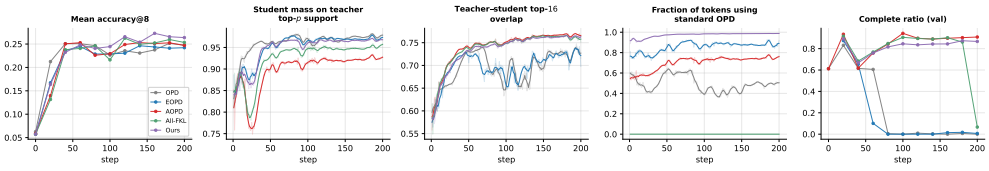


Figure 7: Training dynamics for Qwen3-4B-Base with the Qwen3-8B non-thinking teacher.

Ablation Study. Appendix F presents ablation studies examining sensitivity of our algorithm.

7 Conclusion

Our analysis explains why a strong teacher does not always lead to effective learning under sampled-token OPD. Local updates may provide little correction to teacher-preferred tokens, gradients from other positions can alter update directions, and continued training can lead to repetitive trajectories with high teacher–student overlap. These findings help explain the benefits of SFT initialization and motivate state-dependent KL routing, which improves performance across Qwen3 model scales on mathematical reasoning benchmarks.

AI USE STATEMENT

We used generative AI tools only to assist with writing and plotting. The authors reviewed all AI-assisted outputs and take full responsibility for the content of the manuscript and all reported results.

ETHICS STATEMENT

This work studies the learning dynamics of on-policy distillation and methods for improving student learning from teacher feedback. Although our experiments focus on mathematical reasoning, the proposed method could also be applied to other language-model tasks, where improved model capabilities may introduce risks of misuse.

REPRODUCIBILITY STATEMENT

Our experiments use public datasets and models, with training implemented using the open-source verl framework. The appendix E documents the training hyperparameters, evaluation protocols, baseline implementations, and additional experimental results to support reproducibility.

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.

Mohammadreza Armandpour, Fatih Ilhan, David Harrison, Ajay Jaiswal, Duc NM Hoang, Fartash Faghri, Yizhe Zhang, Minsik Cho, and Mehrdad Farajtabar. Unmasking on-policy distillation: Where it helps, where it hurts, and why. *arXiv preprint arXiv:2605.10889*, 2026.

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

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiacai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes. *arXiv preprint arXiv:2603.25562*, 2026.

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.

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

Nan Jia, Haojin Yang, Xing Ma, Jiesong Lian, Shuailiang Zhang, Weipeng Zhang, Ke Zeng, Xunliang Cai, and Zequn Sun. Asymmetric on-policy distillation: Bridging exploitation and imitation at the token level. *arXiv preprint arXiv:2605.06387*, 2026.

Woogyeol Jin, Taywon Min, Yongjin Yang, Dennis Wei, Yi Zhou, Swanand Ravindra Kadhe, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. *arXiv preprint arXiv:2603.07079*, 2026.

Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 1317–1327, 2016.

Jongwoo Ko, Sara Abdali, Young Jin Kim, Tianyi Chen, and Pashmina Cameron. Scaling reasoning efficiently via relaxed on-policy distillation. *arXiv preprint arXiv:2603.11137*, 2026.

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

Feng Luo, Yu-Neng Chuang, Guanchu Wang, Zicheng Xu, Xiaotian Han, Tianyi Zhang, and Vladimir Braverman. Demystifying opd: Length inflation and stabilization strategies for large language models. *arXiv preprint arXiv:2604.08527*, 2026.

MMA. MAA invitational competitions. `https://maa.org/maa-invitational-competitions/`. Accessed: September 8, 2026.

MMA. 2023 american mathematics competitions. `https://maa.org/student-programs/amc/`, 2023.

Yi Ren and Danica Sutherland. Learning dynamics of llm finetuning. In *International Conference on Learning Representations*, volume 2025, pp. 70523–70563, 2025.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635. JMLR Workshop and Conference Proceedings, 2011.

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.

Yuanyi Wang, Su Lu, Yanggan Gu, Pengkai Wang, Yifan Yang, Zhaoyi Yan, Congkai Xie, Jianmin Wu, and Hongxia Yang. Not all disagreement is learnable: Token teachability in on-policy distillation. *arXiv preprint arXiv:2605.26844*, 2026.

Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, et al. Mimo-v2-flash technical report. *arXiv preprint arXiv:2601.02780*, 2026.

Xingrun Xing, Haoqing Wang, Boyan Gao, Ziheng Li, and Yehui Tang. Trust region on-policy distillation. *arXiv preprint arXiv:2606.01249*, 2026.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026a.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. Tip: Token importance in on-policy distillation. *arXiv preprint arXiv:2604.14084*, 2026b.

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*, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 3: system demonstrations)*, pp. 400–410, 2024.

Siqi Zhu, Xuyan Ye, Hongyu Lu, Weiye Shi, and Ge Liu. The many faces of on-policy distillation: Pitfalls, mechanisms, and fixes. *arXiv preprint arXiv:2605.11182*, 2026.

A Related Work

Knowledge Distillation. Knowledge distillation trains a student using predictions from a teacher (Hinton et al., 2015). For sequence generation, Kim & Rush (2016) introduced sequence-level distillation using teacher-generated responses. The student is trained on teacher-generated responses but generates its own responses at inference time. This mismatch leads to exposure bias (Bengio et al., 2015), with errors accumulating over long sequences. This motivates on-policy distillation, which provides teacher feedback on student-generated prefixes (Agarwal et al., 2024).

On-Policy Distillation. MiniLLM (Gu et al., 2024) develops an on-policy optimization approach for reverse-KL distillation. GKD (Agarwal et al., 2024) supports different divergence objectives and mixtures of fixed data and student-generated sequences. Recent methods adapt supervision at individual token positions. EOPD (Jin et al., 2026) adds FKL where teacher entropy is high, while AOPD (Jia et al., 2026) uses FKL when the sampled-token advantage is non-positive. TrOPD (Xing et al., 2026) uses the sampled-token teacher–student probability ratio to define a trust region, applies FKL to outliers, and adds teacher-prefix guidance. Other methods adjust or select reverse-KL supervision. REOPOLD (Ko et al., 2026) combines reward clipping with entropy-based token selection. TIP (Xu et al., 2026b) studies token importance through student entropy and teacher–student divergence, while Teachability-Aware OPD (Wang et al., 2026) prioritizes disagreement within the student's current support. Our analysis shows why teacher entropy and sampled advantage alone can miss substantial mismatch: a confident teacher may prefer tokens the student rarely samples, and a positive advantage on the sampled token does not rule out mismatch on other tokens. We therefore use teacher-support coverage and underestimation to identify positions where sampled-token reverse-KL updates provide weak correction, and apply FKL to directly supervise teacher-preferred alternatives.

Learning Dynamics of LLM Post-Training. Ren & Sutherland (2025) provide a framework and analyze how fine-tuning updates affect predictions across examples and responses under different post-training algorithms. For OPD, Li et al. (2026) study teacher–student compatibility, alignment among high-probability tokens, and the benefits of teacher-trajectory cold start. They also examine declining teacher supervision quality on longer student prefixes and suggest that cross-position gradient cancellation may limit learning. We build on these perspectives by tracing sampled-token OPD updates through local distributions, generated trajectories, and shared-parameter optimization. Our analysis connects mode suppression and teacher-mode acquisition, identifies spurious agreement on repetitive trajectories, and directly measures cross-position contributions to realized updates. These results help explain the benefits of SFT initialization and guide the choice of positions at which to apply FKL.

B Proof

B.1 Proof of Proposition 1

Fix a state $s \in \mathcal{S}$. Write $\theta_b^k := \theta_{s,b}^k$, and define

$$\Delta\theta_b \;:=\; \theta_b^{k+1} - \theta_b^k, \qquad \Delta\pi(c|s) \;:=\; \pi_\theta^{k+1}(c|s) - \pi_\theta^k(c|s).$$

By assumption, the logits $\{\theta_b\}_{b\in\mathcal{V}}$ at state s are parameterized independently of those at any other state, so we may treat them as the free variables. Taylor-expanding the softmax around θ^k to first order in $\Delta\theta$,

$$\Delta\pi(c|s) \;=\; \sum_{b\in\mathcal{V}} \frac{\partial \pi_\theta^k(c|s)}{\partial \theta_b}\,\Delta\theta_b \;+\; O(\|\Delta\theta\|^2). \tag{18}$$

The softmax Jacobian is

$$\frac{\partial \pi_\theta^k(c|s)}{\partial \theta_b} \;=\; \pi_\theta^k(c|s)\,\big(\mathbf{1}\{c=b\} - \pi_\theta^k(b|s)\big). \tag{19}$$

At the fixed state s, define the token-level advantage

$$A(s,a) \;:=\; \log \pi_T(a|s) - \log \pi_\theta^k(a|s) \tag{20}$$

and hold it fixed at iteration k to form the local surrogate objective

$$J(\theta) \;=\; \mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\big[A(s,a)\big].$$

At $\theta=\theta^k$, the gradient of $J(\theta)$ equals the gradient of the negative reverse KL between $\pi_\theta(\cdot|s)$ and $\pi_T(\cdot|s)$; the two objectives are not identical for general θ. Drawing one sample $u\sim\pi_\theta^k(\cdot|s)$ and treating $A(s,u)$ as a stop-gradient scalar, the unbiased policy-gradient estimator of $\nabla_\theta J(\theta)|_{\theta=\theta^k}$ has logit-coordinate components

$$\frac{\partial \widehat{J}}{\partial \theta_b} \;=\; A(s,u)\,\frac{\partial \log \pi_\theta^k(u|s)}{\partial \theta_b} \;=\; A(s,u)\,\big(\mathbf{1}\{u=b\}-\pi_\theta^k(b|s)\big), \tag{21}$$

where the second equality uses the standard log-softmax derivative $\partial_{\theta_b}\log\pi_\theta^k(u|s) \;=\; \mathbf{1}\{u=b\}-\pi_\theta^k(b|s)$. One gradient ascent step with learning rate η therefore yields

$$\Delta\theta_b \;=\; \eta\,A(s,u)\,\big(\mathbf{1}\{u=b\}-\pi_\theta^k(b|s)\big). \tag{22}$$

Substituting equation 19 and equation 22 into equation 18, the first-order term at token c is

$$\sum_{b\in\mathcal{V}}\frac{\partial\pi_\theta^k(c|s)}{\partial\theta_b}\,\Delta\theta_b \;=\; \eta\,A(s,u)\,\pi_\theta^k(c|s)\sum_{b\in\mathcal{V}}\big(\mathbf{1}\{c=b\}-\pi_\theta^k(b|s)\big)\big(\mathbf{1}\{u=b\}-\pi_\theta^k(b|s)\big).$$

The remainder term in equation 18 is $O(\|\Delta\theta\|^2)=O(\eta^2A(s,u)^2)$, which we absorb into $O(\eta^2)$ for fixed $A(s,u)$. Defining

$$B(c,u;s) \;:=\; \sum_{b\in\mathcal{V}}\big(\mathbf{1}\{c=b\}-\pi_\theta^k(b|s)\big)\big(\mathbf{1}\{u=b\}-\pi_\theta^k(b|s)\big), \tag{23}$$

we obtain

$$\Delta\pi(c|s) \;=\; \eta\,A(s,u)\,\pi_\theta^k(c|s)\,B(c,u;s) \;+\; O(\eta^2), \tag{24}$$

which is the claim of Proposition 1. □

C Distributional Effects of a Sampled-Token OPD Update

Section 3 characterizes the direction and magnitude of token probability change induced by a single sampled-token OPD update under the idealized softmax policy. Here, we provide a more complete characterization of how the same update redistributes probability mass over the full student distribution.

Recall from Proposition 1 that the first-order probability change of any token $c\in\mathcal{V}$ is

$$\Delta\pi(c\mid s)=\eta\,A(s,u)\,\pi_\theta^k(c\mid s)\,B(c,u;s)+O(\eta^2). \tag{25}$$

The first-order probability change depends on three factors: the sampled-token advantage $A(s,u)$, the student probability $\pi_\theta^k(c\mid s)$, and the coupling term $B(c,u;s)$. We denote the first-order probability change by

$$\Delta_1\pi(c\mid s):=\eta A(s,u)\pi_\theta^k(c\mid s)B(c,u;s),\qquad \Delta\pi(c\mid s)=\Delta_1\pi(c\mid s)+O(\eta^2).$$

The direction statements below refer to $\Delta_1\pi$; when this term vanishes, higher-order terms may still produce a nonzero actual change.

For the sampled token $c=u$, the direction and magnitude admit the simple characterization given in Section 3. We therefore focus below on the non-sampled tokens, for which the same sampled-token update induces a richer redistribution over the vocabulary.

C.1 Non-Sampled Token Redistribution

For a non-sampled token $c\neq u$, the coupling term in Eq. equation 25 reduces to

$$B(c,u;s)=\tau(u;s)-\pi_\theta^k(c\mid s), \tag{26}$$

where

$$\tau(u; s) := \left\|\pi_\theta^k(\cdot \mid s)\right\|_2^2 - \pi_\theta^k(u \mid s). \tag{27}$$

We refer to $\tau(u; s)$ as the *watershed*. It is determined by the current student distribution and the sampled token: the term $\pi_\theta^k(u \mid s)$ captures the probability of the sampled token, while $\|\pi_\theta^k(\cdot \mid s)\|_2^2$ captures the concentration of the full distribution.

Substituting Eq. equation 26 into Eq. equation 25 gives

$$\Delta\pi(c \mid s) = \eta A(s, u)\, \pi_\theta^k(c \mid s)\, \left[\tau(u; s) - \pi_\theta^k(c \mid s)\right] + O(\eta^2), \qquad c \neq u. \tag{28}$$

Equation equation 28 shows that the probability change of a non-sampled token depends on the sampled-token advantage, its own student probability, and the full student distribution.

Direction. Because $\pi_\theta^k(c \mid s) > 0$, the first-order direction of a non-sampled token is determined by the sign of $A(s, u)$ and its position relative to the watershed:

$$\operatorname{sign}\left(\Delta_1\pi(c \mid s)\right) = \begin{cases} +\operatorname{sign}\left(A(s, u)\right), & \pi_\theta^k(c \mid s) < \tau(u; s), \\ 0, & \pi_\theta^k(c \mid s) = \tau(u; s), \\ -\operatorname{sign}\left(A(s, u)\right), & \pi_\theta^k(c \mid s) > \tau(u; s), \end{cases} \tag{29}$$

to first order in η.

Thus, the sampled-token advantage provides a common direction signal, while the current student distribution determines whether each non-sampled token follows or opposes that signal. A token lying exactly at the watershed has no first-order probability change.

Magnitude. The corresponding first-order magnitude is

$$|\Delta\pi(c \mid s)| = \eta\, |A(s, u)|\, \pi_\theta^k(c \mid s)\, \left|\tau(u; s) - \pi_\theta^k(c \mid s)\right| + O(\eta^2). \tag{30}$$

A large sampled-token advantage does not necessarily produce a large probability change for every token. The first-order change also depends on the token's student probability and its distance from the watershed. It is small when the student probability is near zero or close to the watershed.

The watershed therefore controls both aspects of the non-sampled dynamics: crossing the watershed reverses the first-order direction, while approaching the watershed reduces the first-order magnitude.

C.2 Geometry of the Redistribution

We next characterize the location of the watershed relative to the current student distribution. This geometry determines how the sampled-token advantage affects the probabilities of non-sampled tokens.

Let

$$c^\star := \arg\max_{b \in \mathcal{V}} \pi_\theta^k(b \mid s) \tag{31}$$

denote the highest-probability student token.

Theorem 1 (Mode-direction invariance)**.** *For every sampled token $u \in \mathcal{V}$,*

$$\pi_\theta^k(c^\star \mid s) > \tau(u; s). \tag{32}$$

Consequently, whenever $c^\star \neq u$,

$$\operatorname{sign}\left(\Delta_1\pi(c^\star \mid s)\right) = -\operatorname{sign}\left(A(s, u)\right) \tag{33}$$

to first order in η.

Proof. Since

$$\pi_\theta^k(b \mid s) \leq \pi_\theta^k(c^\star \mid s) \qquad \text{for every } b \in \mathcal{V},$$

we have

$$\left\|\pi_\theta^k(\cdot \mid s)\right\|_2^2 = \sum_{b \in \mathcal{V}} \pi_\theta^k(b \mid s)^2 \leq \pi_\theta^k(c^\star \mid s) \sum_{b \in \mathcal{V}} \pi_\theta^k(b \mid s) = \pi_\theta^k(c^\star \mid s). \tag{34}$$

Therefore,

$$\tau(u;s) = \left\|\pi_\theta^k(\cdot \mid s)\right\|_2^2 - \pi_\theta^k(u \mid s) \le \pi_\theta^k(c^\star \mid s) - \pi_\theta^k(u \mid s) < \pi_\theta^k(c^\star \mid s), \tag{35}$$

where the strict inequality follows from $\pi_\theta^k(u \mid s) > 0$ under the softmax parameterization. □

Theorem 1 shows that whenever the student mode is not itself sampled, it always lies above the watershed and therefore moves against the sampled-token advantage.

Sampling the student mode. The redistribution simplifies further when the sampled token is itself the student mode.

Corollary 1 (Mode sampling)**.** *If $u = c^\star$, then*

$$\tau(u;s) \le 0. \tag{36}$$

Consequently, every non-sampled token satisfies

$$\operatorname{sign}\left(\Delta_1 \pi(c \mid s)\right) = -\operatorname{sign}\left(A(s,u)\right), \qquad c \neq u, \tag{37}$$

to first order in η.

Proof. If $u = c^\star$, then

$$\pi_\theta^k(b \mid s) \le \pi_\theta^k(u \mid s) \qquad \text{for every } b \in \mathcal{V}.$$

Hence,

$$\left\|\pi_\theta^k(\cdot \mid s)\right\|_2^2 = \sum_{b \in \mathcal{V}} \pi_\theta^k(b \mid s)^2 \le \pi_\theta^k(u \mid s) \sum_{b \in \mathcal{V}} \pi_\theta^k(b \mid s) = \pi_\theta^k(u \mid s), \tag{38}$$

and therefore

$$\tau(u;s) = \left\|\pi_\theta^k(\cdot \mid s)\right\|_2^2 - \pi_\theta^k(u \mid s) \le 0. \tag{39}$$

Since every non-sampled softmax probability is strictly positive,

$$\pi_\theta^k(c \mid s) > \tau(u;s) \qquad \text{for every } c \neq u,$$

and the result follows from Eq. equation 29. □

Thus, mode sampling induces a particularly simple distribution-wide redistribution: the sampled mode moves with its advantage, while all remaining tokens move in the opposite direction.

Redistribution regimes. More generally, the sign of the watershed distinguishes two qualitatively different redistribution regimes.

Non-positive watershed: $\tau(u;s) \le 0$. Since every softmax probability is strictly positive,

$$\pi_\theta^k(c \mid s) > \tau(u;s), \qquad c \neq u. \tag{40}$$

Hence all non-sampled tokens move against the sampled-token advantage:

$$\operatorname{sign}\left(\Delta_1 \pi(c \mid s)\right) = -\operatorname{sign}\left(A(s,u)\right), \qquad c \neq u, \tag{41}$$

to first order. Mode sampling always belongs to this regime, although a non-positive watershed can also occur when an off-mode token is sampled.

Positive watershed: $\tau(u;s) > 0$. Non-sampled tokens may lie on either side of the watershed. Tokens satisfying

$$\pi_\theta^k(c \mid s) < \tau(u;s) \tag{42}$$

move with the sampled-token advantage, whereas tokens satisfying

$$\pi_\theta^k(c \mid s) > \tau(u;s) \tag{43}$$

move against it. Tokens satisfying

$$\pi_\theta^k(c \mid s) = \tau(u;s) \tag{44}$$

have zero first-order probability change.

Thus, when the watershed is positive, non-sampled token probabilities can change in different directions. However, a given distribution need not have non-sampled tokens on both sides of the watershed.

C.3 Probability Changes Relative to the Teacher

The preceding analysis characterizes whether a non-sampled token moves with or against the *sampled-token advantage*. This is distinct from asking whether that token moves toward its own probability under the teacher.

For each token c, define the teacher-preferred direction as

$$D_T(c; s) := \operatorname{sign}\left(\pi_T(c \mid s) - \pi_\theta^k(c \mid s)\right). \tag{45}$$

For a token whose student and teacher probabilities differ, the first-order change points toward the teacher when

$$\operatorname{sign}\left(\Delta_1 \pi(c \mid s)\right) D_T(c; s) > 0, \tag{46}$$

and away from the teacher when this product is negative. If $\Delta_1 \pi(c \mid s) = 0$, there is no first-order movement. If $D_T(c; s) = 0$, the probabilities already match, and any nonzero first-order change moves the student away from that match.

For the sampled token u, these two notions coincide. Since the logarithm is monotone,

$$\operatorname{sign}\left(A(s, u)\right) = \operatorname{sign}\left(\pi_T(u \mid s) - \pi_\theta^k(u \mid s)\right). \tag{47}$$

Together with the sampled-token direction established in Section 3, this implies that the sampled token moves toward its corresponding teacher probability to first order when the probabilities differ, and remains unchanged when they match.

For a non-sampled token $c \neq u$, however, the first-order update direction is

$$\operatorname{sign}\left(\Delta_1 \pi(c \mid s)\right) = \operatorname{sign}\left(A(s, u)\right) \operatorname{sign}\left(\tau(u; s) - \pi_\theta^k(c \mid s)\right), \tag{48}$$

which need not coincide with $D_T(c; s)$. Equivalently, a non-sampled token is moved toward the teacher when

$$\operatorname{sign}\left(A(s, u)\right) \operatorname{sign}\left(\tau(u; s) - \pi_\theta^k(c \mid s)\right) D_T(c; s) > 0, \tag{49}$$

and away from the teacher when the product is negative.

Equation equation 49 also gives a geometric interpretation in the student–teacher probability plane. The diagonal

$$\pi_T(c \mid s) = \pi_\theta^k(c \mid s) \tag{50}$$

separates the two teacher-preferred directions, while the vertical boundary

$$\pi_\theta^k(c \mid s) = \tau(u; s) \tag{51}$$

separates the two directions induced by the sampled-token update. Together, these boundaries partition the plane into regions where the induced probability change is either aligned with or opposed to the teacher.

Consequently, a single sampled-token OPD step need not move every token toward its corresponding teacher probability. The sampled token receives a teacher-directed correction, whereas the remaining tokens respond indirectly through the redistribution induced by the sampled-token advantage and the current student distribution.

C.4 Empirical Validation of Proposition 1

In this subsection, we validate the correctness of Proposition 1 using a categorical policy with $|\mathcal{V}| = 100$. We randomly initialize student and teacher logits and obtain their categorical distributions through softmax. For each trial, we sample $u \sim \pi_\theta(\cdot \mid s)$ and apply one SGD update.

For each update, we compute the predicted changes in log probability and probability from Proposition 1, using the pre-update distributions. We measure the actual changes by subtracting the pre-update values from the post-update values, as defined in equation 5, and compare the predictions with these measurements across all tokens and random initializations. The close agreement with the identity line validates the per-token dynamics in Proposition 1.

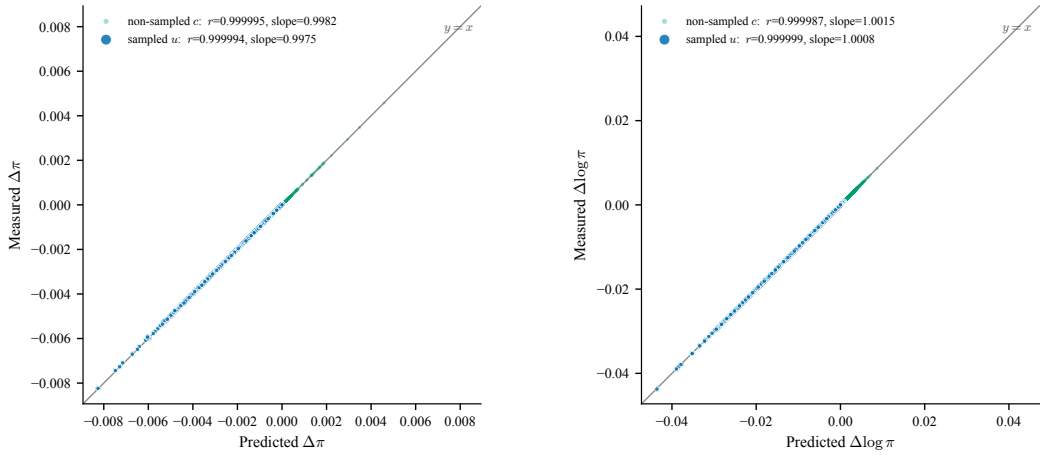


Figure 8: **Validation of Proposition 1 under randomly initialized categorical policies.** Predicted and measured changes are shown for $\Delta \log \pi$ and $\Delta\pi$. Each point represents one vocabulary token after a single sampled-token update; the dashed line denotes $y = x$.

C.5 Controlled Token-Level Interventions

The preceding analysis separates the direction induced by the sampled-token update from the direction preferred by the teacher. We next isolate the sources of this behavior through two controlled interventions. First, we vary the sampled token while holding both the student and teacher distributions fixed. We then vary the teacher distribution while holding the student distribution and sampled token fixed.

Varying the sampled token under fixed student and teacher distributions. We first hold $\pi_\theta^k(\cdot \mid s)$ and $\pi_T(\cdot \mid s)$ fixed and vary only the sampled token u. This isolates the effect of the sampled token on the one-step redistribution.

Changing u modifies the update through two quantities. First, it changes the sampled-token advantage

$$A(s, u) = \log \pi_T(u \mid s) - \log \pi_\theta^k(u \mid s). \tag{52}$$

Second, it changes the watershed

$$\tau(u; s) = \left\|\pi_\theta^k(\cdot \mid s)\right\|_2^2 - \pi_\theta^k(u \mid s). \tag{53}$$

Thus, even with the same teacher and student distributions, changing the sampled token can change both the advantage $A(s, u)$ and the watershed $\tau(u; s)$.

Figure 9 shows that the identity of the sampled token can qualitatively change the resulting redistribution. For example, the same token can move away from its teacher probability under one sampled token but toward the teacher under another. In the illustrated cases, token 35 exhibits such a change despite both the teacher and student distributions remaining fixed.

The difference arises entirely from the sampled-token-dependent terms in Eq. equation 28: changing u changes both the sign and magnitude of $A(s, u)$ and the location of the watershed $\tau(u; s)$. Thus, the teacher–student pair alone does not determine the one-step probability change of the full distribution; the sampled token is an essential part of the dynamics.

Varying the teacher while fixing the student distribution and sampled token. We next fix the student distribution and sampled token u, while varying only the teacher distribution. Since $\pi_\theta^k(\cdot \mid s)$ and u are unchanged, the student-side watershed

$$\tau(u; s) = \left\|\pi_\theta^k(\cdot \mid s)\right\|_2^2 - \pi_\theta^k(u \mid s) \tag{54}$$

is identical across all cases. The student probability $\pi_\theta^k(c \mid s)$ of each token c is also unchanged.

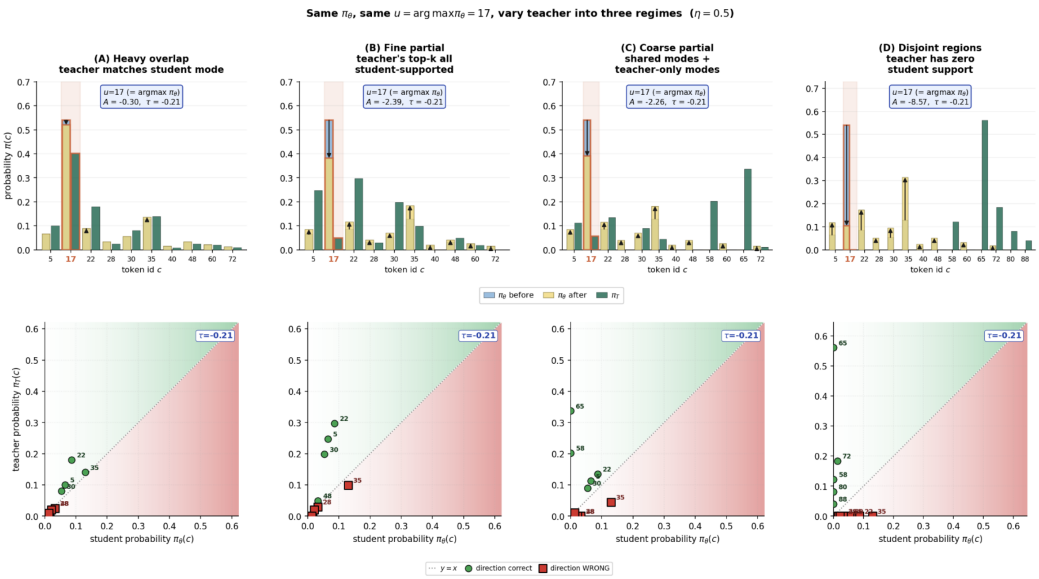


Figure 9: **Changing the sampled token changes the one-step redistribution under fixed student and teacher distributions.** Across all columns, $\pi_\theta^k(\cdot \mid s)$ and $\pi_T(\cdot \mid s)$ are fixed; only the sampled token u changes. *Top:* the student distribution before and after one OPD update together with the teacher distribution. The sampled token is highlighted. *Bottom:* each token is plotted according to its student and teacher probabilities. The diagonal denotes equal student and teacher probability, while the vertical line marks the watershed $\tau(u; s)$. Green and red regions indicate probability changes toward and away from the teacher, respectively. Changing the sampled token modifies both $A(s, u)$ and $\tau(u; s)$, and can therefore change the teacher alignment of the same token even though the teacher and student distributions are unchanged.

Changing the teacher nevertheless affects the dynamics in two ways. It changes the sampled-token advantage $A(s, u)$, which controls the common sign and scale of the induced redistribution, and it changes the teacher-preferred direction $D_T(c; s)$ of each token. This intervention therefore holds the student-side geometry fixed while varying the structure of the teacher–student mismatch.

We consider teacher distributions with progressively less probability-mass overlap with the student.

Figure 10 illustrates several qualitative behaviors. When teacher and student probability mass substantially overlap, much of the induced redistribution can align with the teacher. Under partial overlap, however, individual non-sampled tokens can move in conflict with their teacher-preferred directions; token 35 provides one such example.

As the teacher and student distributions overlap less, tokens favored by the teacher but assigned very low student probabilities may move toward their teacher probabilities only slightly. Meanwhile, tokens with higher student probabilities may move farther away from their teacher probabilities.

This asymmetry follows directly from the magnitude of the probability change for a non-sampled token

$$|\Delta\pi(c \mid s)| = \eta\,|A(s, u)|\,\pi_\theta^k(c \mid s)\,\big|\tau(u; s) - \pi_\theta^k(c \mid s)\big| + O(\eta^2). \tag{55}$$

Even when the induced direction is teacher-aligned, a token with near-zero student probability receives little probability change because of the multiplicative factor $\pi_\theta^k(c \mid s)$. Conversely, a higher-probability token can undergo a substantially larger probability change even when that probability change is opposed to the teacher.

These two experiments separate the effects of the sampled token and the teacher distribution. Changing the sampled token changes both $A(s, u)$ and $\tau(u; s)$. Changing the teacher while fixing the student distribution and sampled token leaves $\tau(u; s)$ unchanged, but changes $A(s, u)$ and whether each token's probability change is toward or away from the teacher.

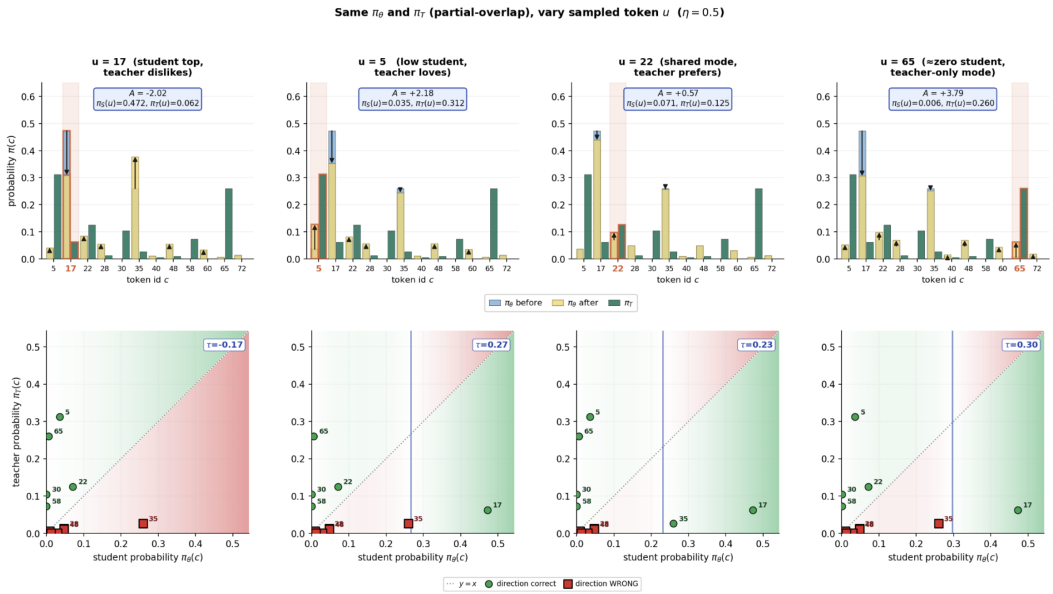


Figure 10: **Teacher alignment changes with the mismatch structure under fixed student-side geometry.** Across all columns, the student distribution and sampled token u are fixed, while the teacher distribution has progressively less probability-mass overlap with the student. Consequently, the watershed $\tau(u; s)$ and the student token probabilities remain unchanged, whereas the sampled-token advantage $A(s, u)$ and the teacher-preferred directions vary. *Top:* the student distribution before and after one OPD update together with each teacher distribution. *Bottom:* token-level student and teacher probabilities, with green and red regions indicating probability changes toward and away from the teacher, respectively. The same student-side probability geometry can therefore produce substantially different patterns of teacher alignment as the mismatch structure changes.

These examples make explicit the factorization in Eq. equation 25: the probability change of a token is jointly determined by the sampled-token advantage, its own current probability, and the geometry of the student distribution. Whether that probability change is toward or away from the teacher is a separate property determined by how the induced direction aligns with the teacher–student mismatch at that token.

D Diagnostic Experiments: Setup and Additional Results

D.1 Experimental Setup

We use several teacher–student pairs to examine how OPD training dynamics vary across student and teacher models. Table 2 defines the abbreviations used throughout the diagnostic figures. We construct the held-out prefixes from responses generated by the teacher model on the AIME dataset.

Overlap and overlap mass. At a state s, let $\mathcal{S}_K^\theta(s)$ denote the student's top-K token set and let $\mathcal{S}_K(s)$ denote the teacher's top-K token set, as defined in the main text. Their overlap is the intersection

$$\mathcal{I}_K(s) = \mathcal{S}_K^\theta(s) \cap \mathcal{S}_K(s).$$

The student and teacher probability masses on this intersection are

$$M^\theta_{\text{overlap}}(s) = \sum_{c \in \mathcal{I}_K(s)} \pi_\theta(c \mid s), \qquad M^T_{\text{overlap}}(s) = \sum_{c \in \mathcal{I}_K(s)} \pi_T(c \mid s).$$

We use $K = 16$ for these overlap diagnostics. Both masses use the original probabilities without renormalization. Unlike student coverage of the teacher's top-K set, overlap mass sums only over tokens present in both top-K sets.

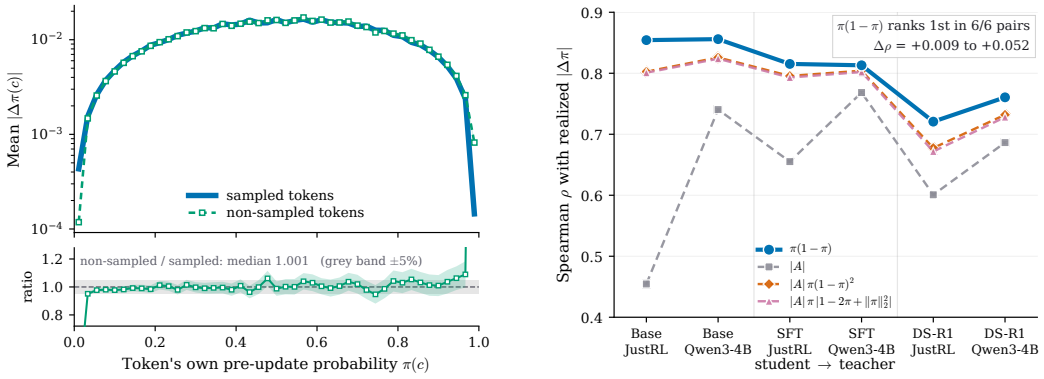
Abbreviation	Model or checkpoint
Qwen3-1.7B-Base	`Qwen3-1.7B-Base`
Qwen3-1.7B-SFT	SFT of Qwen3-1.7B-Base on Qwen3-4B
R1-Distill-1.5B	`DeepSeek-R1-Distill-Qwen-1.5B`
JustRL-1.5B	`hbx/JustRL-DeepSeek-1.5B`
Qwen3-4B	`Qwen3-4B`
Qwen3-8B	`Qwen3-8B`

Table 2: Model abbreviations used in the diagnostic experiments.

D.2 Additional Results on Local Distributional Dynamics

We further examine the relationship between student probability and realized token probability change, measured at the same prefixes before and after OPD training.

Figure 11(a) shows that sampled and non-sampled tokens exhibit similar magnitudes of change $|\Delta\pi|$ at comparable student probabilities. Non-sampled tokens are selected from the union of the student and teacher top-K sets. Figure 11(b) shows that probability-based measures predict $|\Delta\pi|$ more accurately than the advantage-based measures considered. Therefore, these results show that realized probability change depends on the student's current distribution, beyond the sampled-token advantage alone.



(a) Sampled vs. non-sampled tokens.

(b) Predictor comparison.

Figure 11: **Additional comparisons of token probability change under OPD. (a)** Probability change for sampled and non-sampled tokens, grouped by pre-update student probability. **(b)** Predictive performance of probability-based and advantage-based quantities for realized $|\Delta\pi|$.

D.3 Validation of the Cross-Position Decomposition

We further examine whether sign agreement varies with the magnitude of the advantage or the student's pre-update probability. As shown in Figure 12, sign agreement remains near 50% across both factors, with no clear dependence on either.

Predicting the log-probability change from row sums. Under the first-order approximation, the realized change at receiver position i is

$$\Delta \log \pi_S(u_i) = \sum_j M_{ij} + O(\eta^2).$$

As shown in Fig. 14(a), the row sum $\widehat{\Delta}_i = \sum_j M_{ij}$ closely matches the measured one-step change in sampled-token log probability after an actual SGD update. This confirms that the interaction matrix captures the realized joint update to first order.

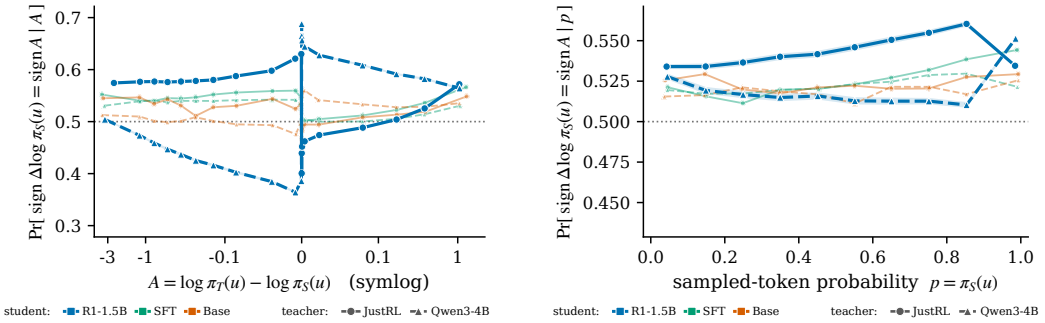


Figure 12: Sign agreement between the student's probability change and the advantage, conditioned on (a) advantage magnitude and (b) the student's pre-update probability. Sign agreement remains near 50% across both factors.

Residual scaling with step size. We repeat the measurement at different learning rates to check whether the prediction error follows the expected $O(\eta^2)$ scaling. Fig. 14(b) shows that the residual between $\widehat{\Delta}_i$ and the measured update decreases as $O(\eta^2)$. The observed scaling is therefore consistent with the first-order Taylor approximation underlying the decomposition.

Cross-position interactions depend on both the student and the teacher. Having shown that cross-position contributions can override a token's own update, we examine how these contributions vary across teacher–student pairs. Figure 13 shows substantial off-diagonal entries in all four settings, with different patterns and magnitudes across Base and SFT students and their teachers. These differences reflect the two factors in $M_{ij} = \eta A_j \langle h_i, h_j \rangle$: the student's gradient alignment determines how positions are coupled, while the teacher-dependent advantage weights each sender's contribution. Thus, the realized update at a position depends on both its coupling to other positions and the supervision those positions receive.

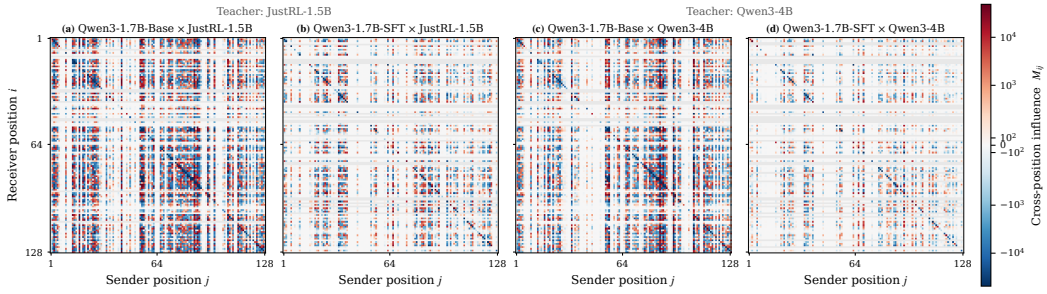


Figure 13: **Cross-position interaction structure depends on both the student state and the teacher.** Each panel shows the exact interaction matrix M_{ij} for one teacher–student pair, with sender position j on the x-axis and receiver position i on the y-axis. Positive and negative influences are shown on a shared symmetric scale.

Leave-one-out validation of individual interactions. The row-sum result validates the aggregate decomposition but does not by itself establish that an individual entry M_{ij} can be interpreted as the influence of sender j on receiver i. We therefore compare M_{ij} with a leave-one-out intervention in which sender j is removed from the update and the resulting change at receiver i is measured directly.

As shown in Fig. 14(c), interactions large enough to measure reliably closely match their leave-one-out effects. For the strongest 20% of interactions by $|M_{ij}|$, we obtain Pearson $r = 0.941$, slope $= 0.987$, and sign agreement $= 0.924$. Weaker interactions are too small to measure reliably given

Step 10

Rank	Token	A	Weight
1	U+0627 U+0643	−40.38	57.30%
2	paque	−9.00	17.88%
3	tens	−5.14	4.32%
4	<|endoftext|>	−21.16	2.83%
5	a	−1.41	1.96%
6	U+653F U+5E9C	−1.99	1.67%
7	written	1.74	1.44%
8	The	−3.04	1.28%
9	at	−1.79	1.16%
10	$	−1.00	1.08%

Step 50

Rank	Token	A	Weight
1	But	−2.81	12.05%
2	\(	−6.07	8.10%
3	pairwise	−5.06	7.78%
4	,	−3.19	5.20%
5	dots	−8.99	4.95%
6	pair	1.39	4.92%
7	But	−2.91	4.64%
8	positive	−2.18	4.33%
9	circle	−1.44	3.70%
10	7	1.78	3.08%

Step 100

Rank	Token	A	Weight
1	<|endoftext|>	−22.53	31.81%
2	Rightarrow	3.76	10.09%
3	that	2.09	6.20%
4	become	−2.61	3.55%
5	terms	−1.35	3.03%
6	value	−1.96	2.71%
7	which	−1.52	2.59%
8	2	−0.89	2.16%
9	let	−0.85	2.07%
10	(	−1.39	2.02%

Table 3: Top ten tokens by absolute cross-position influence $\sum_{i \neq j} |M_{ij}|$ at three checkpoints. Ranks and weights are computed separately within each checkpoint's 200 selected positions, including one explicitly included termination position. Repeated token strings denote different positions. Non-Latin tokens are represented by their Unicode code points.

the numerical precision of the forward pass. These results support interpreting M_{ij} as the first-order sender-to-receiver influence of position j on position i.

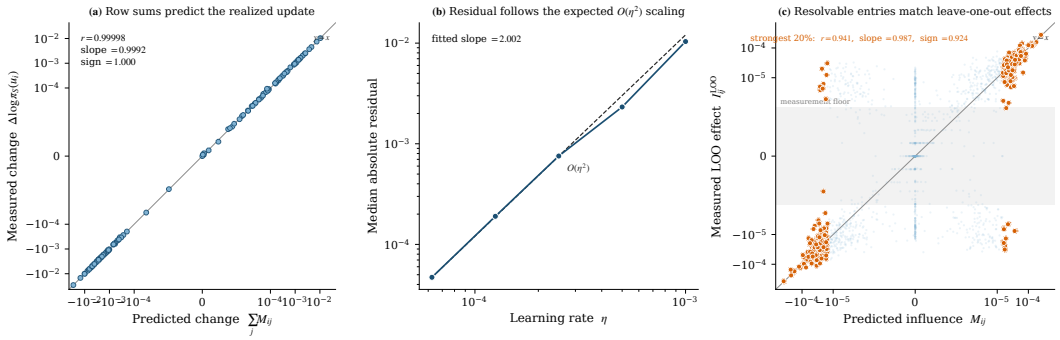


Figure 14: **Validation of the cross-position influence decomposition. (a)** The row sum $\sum_j M_{ij}$ closely matches the measured one-step change $\Delta \log \pi_S(u_i)$. **(b)** The approximation residual decreases as $O(\eta^2)$ with the learning rate, consistent with a first-order expansion. **(c)** Individual M_{ij} values agree with leave-one-out marginal effects when the interaction magnitude is above the numerical measurement floor; the strongest 20% by $|M_{ij}|$ are highlighted.

E Experimental Settings

Models and Training Datasets. We conduct experiments using three Qwen3 (Yang et al., 2025) student models of different sizes, Qwen3-1.7B-Base, Qwen3-4B-Base, and Qwen3-8B-Base. We use Qwen3-8B as the teacher model in both settings: non-thinking mode for OPD directly from Base initialization, and thinking mode for OPD from SFT initialization. We use the DAPO dataset (Yu et al., 2026) for OPD training across all student sizes. For SFT, we extract 400K prompt–response pairs from `open-thoughts/OpenThoughts3-1.2M`, comprising 50K distinct prompts with eight responses each.

Evaluation Benchmarks and Metrics. We evaluate on AIME24-26 (MMA) and AMC23 (MMA, 2023). We generate eight responses per question using different validation settings for the Base and SFT-initialized cases, as specified in Table 4. We report Avg@8, the mean accuracy across the eight responses, and Best@8, estimated by bootstrap resampling. For each question, we repeatedly draw eight responses with replacement from its eight generated responses and record whether any resampled response is correct. We average these binary outcomes over bootstrap replicates and then over questions. If c_i of the eight original responses to question i are correct, the conditional expectation of its bootstrap score is $1 - (1 - c_i/8)^8$. Thus, for N questions, the expected bootstrap

metric is

$$\text{Best@8} = \frac{1}{N}\sum_{i=1}^{N}\left[1 - \left(1 - \frac{c_i}{8}\right)^{8}\right].$$

Table 1 reports the bootstrap estimates as percentages. This metric differs from the fraction of questions with at least one correct response among the original eight generations.

Baselines. We compare our design against three baselines: (i) OPD (Lu & Lab, 2025), standard on-policy distillation with sampled-token reverse KL; (ii) EOPD (Jin et al., 2026), which uses forward KL when the teacher has high entropy; (iii) AOPD (Jia et al., 2026), which uses FKL when the sampled-token advantage is non-positive. For the training-dynamics comparison in Figure 7, we additionally include All-FKL, which applies the same top-K FKL term as in equation 17 at every token position, corresponding to $r(s_t) = 1$ for all t. At fixed K, All-FKL requires more teacher-target storage than selective routing when targets are retained only at positions receiving FKL. This storage comparison does not establish a reduction in peak training memory, which also depends on the implementation. Jin et al. (2026) discuss the memory overhead of full-vocabulary FKL and the linear growth of top-K target storage with K.

Implementation We implement on-policy distillation with verl. We use a training batch size of 128 and a mini-batch size of 128 to ensure an on-policy setting. The main hyperparameters are summarized in Table 4. For EOPD (Jin et al., 2026), we use a teacher entropy threshold $\tau = 1.0$; for AOPD, we use an advantage threshold of 0 to apply FKL.

Hyperparameter	**OPD**	
	Base Model	**SFT Model**
Training steps	200	100
Train batch size	128	128
Micro batch size	128	128
Rollout n	1	1
Rollout Correction	[0.5, 2.0]	[0.5, 2.0]
Maximum prompt length	2048	2048
Maximum response length	4096	16384
Validation max response length	8192	16384
Learning rate	1×10^{-6}	1×10^{-6}
Temperature	1.0	1.0
Top-p	1.0	1.0
Validation temperature	1.0	0.6
Validation top-p	0.8	0.95
Validation top-k	-1	20
Loss aggregation	`token-mean`	`token-mean`

Table 4: Default hyperparameters for OPD. For state-dependent KL routing, we use $p = 0.9$ for the teacher top-p set, $K = 16$ for approximate top-K FKL, and thresholds $\epsilon = 6$, $\epsilon_2 = 0.3$, and $\epsilon_3 = 0.2$ in equation 16–equation 17.

For SFT, we fine-tune all model parameters on the extracted 400K prompt–response pairs using LLaMA-Factory (Zheng et al., 2024). The eight responses per prompt are taken directly from `OpenThoughts3-1.2M`. The prompt template is:

Prompt template

{Question} Please reason step by step, and put your final answer within `\boxed{}`.

We format the responses to match the template used by Qwen3-8B in thinking mode, ensuring consistent formatting between SFT and subsequent OPD training. The main SFT hyperparameters are summarized in Table 5.

Hyperparameter	Value
Student model	Base model
Training objective	Full-parameter SFT
Template	`qwen3`
Training epochs	1
Sequence length	20,000
Per-device batch size	8
Gradient accumulation steps	1
Batch Size	512
Learning rate	8×10^{-5}
LR scheduler	Cosine
Warmup ratio	0.1
Precision	BF16

Table 5: SFT hyperparameters for cold-start initialization of the 1.7B, 4B, and 8B base models using `OpenThoughts3-1.2M`.

F Ablation Study

We examine the routing thresholds, candidate set, and individual routing rules using Qwen3-1.7B-Base with the Qwen3-8B teacher. Unless otherwise specified, we use $\varepsilon_1 = 6$, $\varepsilon_2 = 0.3$, $\varepsilon_3 = 0.2$, and $p = 0.90$.

Routing thresholds. Table 6 shows that increasing ε_1 from 1.5 to 6 improves mean Avg@8 from 16.15% to 18.12%, while reducing the final FKL fraction from 17.6% to 6.7%. Further reducing FKL use does not necessarily improve performance: increasing ε_2 from 0.3 to 0.5 lowers the FKL fraction to 3.6%, but also reduces mean Avg@8 to 16.90%. Among the tested coverage thresholds, $\varepsilon_3 = 0.2$ achieves the highest mean Avg@8. These results show that the choice of routing thresholds matters beyond the amount of FKL supervision.

Parameter	Value	f_{FKL} (%)	AIME24	AIME25	AIME26	AMC23	Avg.
ε_1	1.5	17.6	9.17	7.50	7.92	40.00	16.15
	3	6.7	10.42	11.67	6.25	39.38	16.93
	6	6.7	12.08	9.58	5.83	45.00	**18.12**
ε_2	0.2	9.8	10.00	5.83	6.67	40.94	15.86
	0.3	6.7	12.08	9.58	5.83	45.00	**18.12**
	0.5	3.6	9.17	8.75	7.50	42.19	16.90
	0.7	1.8	10.00	5.83	7.92	39.69	15.86
ε_3	0.1	6.9	12.50	6.67	6.67	36.56	15.60
	0.2	6.7	12.08	9.58	5.83	45.00	**18.12**
	0.7	8.7	11.67	5.00	7.92	43.12	16.93

Table 6: Parameter sensitivity on Qwen3-1.7B-Base with the Qwen3-8B teacher. We vary one threshold at a time, holding the others at $\varepsilon_1 = 6$, $\varepsilon_2 = 0.3$, $\varepsilon_3 = 0.2$, and $p = 0.90$. Scores are Avg@8 (%) at each run's best validation checkpoint over 200 training steps. f_{FKL} is the fraction of valid response positions routed to FKL at the final step.

Comparison between $M_{\mathrm{top-p}}$ and $M_{\mathrm{top-k}}$ on underestimated token In Table 7, we compare using the student mass on teacher's adaptive top-p set and top-k set to estimate the underestimated token for $V(\theta; s_t)$. Thus, using student mass on adaptive teacher top-p set for routing is more helpful compared with on top-K set , consistent with our local distributional analysis.

Routing rules. Table 8 compares the coverage rule $M_{\text{top-}p}$, the underestimation vote V, and their combination. Combining both rules achieves the highest mean Avg@8 of 18.12%, compared with 17.06% for coverage alone and 16.54% for the vote alone. The combined rule and vote-only variant have the same final FKL fraction of 6.7%, yet differ by 1.58 accuracy points. These results suggest that the two rules provide complementary criteria for selecting positions for FKL.

Candidate set	f_{FKL} (%)	AIME24	AIME25	AIME26	AMC23	Avg.
Top-p, $p = 0.90$	6.7	12.08	9.58	5.83	45.00	**18.12**
Top-p, $p = 0.95$	7.1	12.92	5.83	7.92	40.94	16.90
Fixed top-K	6.8	11.25	6.25	6.67	40.31	16.12

Table 7: Comparison between using $M_{\mathrm{top-p}}$ and $M_{\mathrm{top-k}}$ for mismatch states.

Active rules	f_{FKL} (%)	AIME24	AIME25	AIME26	AMC23	Avg.
$M_{\mathrm{top-p}}$ only	0.4	12.08	7.08	8.75	40.31	17.06
V only	6.7	10.00	6.67	7.92	41.56	16.54
$M_{\mathrm{top-p}}$ + V	6.7	12.08	9.58	5.83	45.00	**18.12**

Table 8: Effect of the active routing rules.