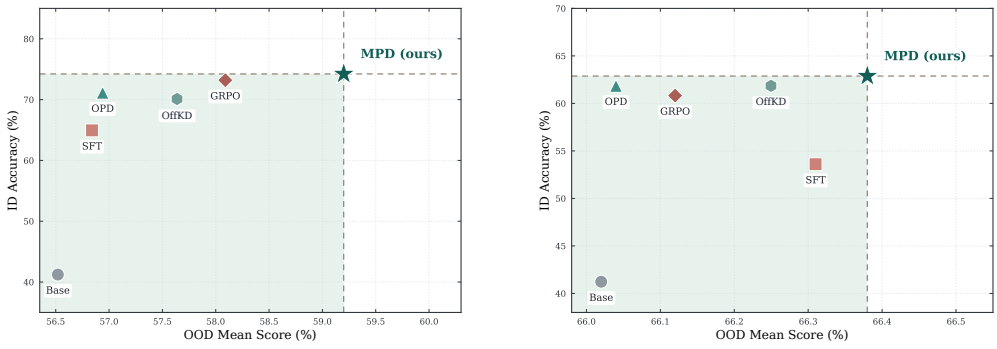
MIXED-POLICY DISTILLATION: BRINGING TEACHER AND STUDENT ONTO COMMON GROUND

Anonymous authors
Paper under double-blind review

ABSTRACT

Distillation involves both the student and teacher policies, but trajectories aligned with one are not necessarily suitable for the other. In language model post-training, teacher-generated trajectories can depart from the student's inference distribution, while on-policy distillation (OPD) can expose the teacher to unfamiliar student-generated states and weaken its guidance. We introduce Mixed-Policy Distillation (MPD), a post-training framework that constructs trajectories for compatibility with both teacher and student. MPD combines a student-generated prefix with a teacher-generated continuation through an uncertainty-guided handoff and distills the teacher's token distributions across the full mixed trajectory. In this design, the teacher helps shape subsequent states while its guidance remains anchored in the student's own reasoning. Empirical analysis shows that, after the handoff, mixed trajectories achieve lower joint perplexity and stronger next-token agreement than trajectories generated entirely by either policy alone. Experiments across mathematics, science, and tool use on various Qwen and Gemma models demonstrate improved performance over both off-policy and on-policy counterparts with both externally trained teachers and privileged-information-conditioned self-teachers. These results identify mixed-policy trajectory construction as a promising design choice for distillation-based post-training.



(a) Qwen3.5-4B (b) Gemma3-4B

Figure 1: In-domain (ID) and out-of-domain (OOD) performance in the tool use experiments on (a) Qwen3.5-4B and (b) Gemma3-4B. The vertical axis reports ToolAlpaca accuracy, and the horizontal axis reports mean score across six general-domain benchmarks; higher is better on both axes. OffKD (off-policy knowledge distillation; Section 2), OPD, and MPD use the same GRPO-trained teacher. Dashed lines indicate MPD's scores. MPD attains the highest ID and OOD performance on both models.

1 INTRODUCTION

On-policy distillation (OPD) has emerged as a widely used approach to language-model post-training, in which a student learns from a teacher's token-level distributions along student-generated trajectories (Agarwal et al., 2024). Compared with supervised fine-tuning (SFT) on teacher-generated demonstrations, OPD has been shown to induce less distributional shift from the base

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.

model and better preserve its performance on general-domain tasks (Agarwal et al., 2024). OPD also extends to on-policy self-distillation (OPSD), which replaces the external teacher with a privileged-information-conditioned self-teacher (Zhao et al., 2026; Shenfeld et al., 2026). OPD and related variants have been adopted in recent foundation models, including DeepSeek-V4 (Xu et al., 2026), Qwen3 (Yang et al., 2025), GLM-5 (Zeng et al., 2026), and Kimi K3 (Team et al., 2026).

Despite its broad adoption, OPD faces a central limitation: student-generated trajectories are not necessarily suitable for teacher supervision. The resulting histories can enter low-probability regions under the teacher, where its conditional distributions may provide less reliable learning targets and insufficient corrective signal in degenerate cases such as endless repetition or severely erroneous reasoning paths (Xu et al., 2025; Liu et al., 2026; Jiang et al., 2026; Zhao et al., 2026). Conversely, teacher-generated trajectories introduce a complementary problem: although well aligned with the teacher, they may deviate from the student's inference distribution and provide supervision at prefixes that the student is unlikely to encounter (Kim & Rush, 2016; Wu et al., 2026; Zhang et al., 2026b). These opposing mismatches motivate constructing trajectories for joint teacher–student compatibility rather than drawing them from either policy alone.

To address the problem, we propose Mixed-Policy Distillation (MPD), which constructs each training trajectory from a student-generated prefix and a teacher-generated continuation. MPD first samples a student rollout and selects a handoff point using an entropy-guided strategy. Specifically, it chooses the position with the highest next-token entropy, where the student is most uncertain about how to continue. MPD retains the student-generated prefix up to the handoff point, and the teacher conditions on this prefix to generate the remaining continuation. The student is then trained on the complete mixed trajectory by matching the teacher's next-token distribution at every position.

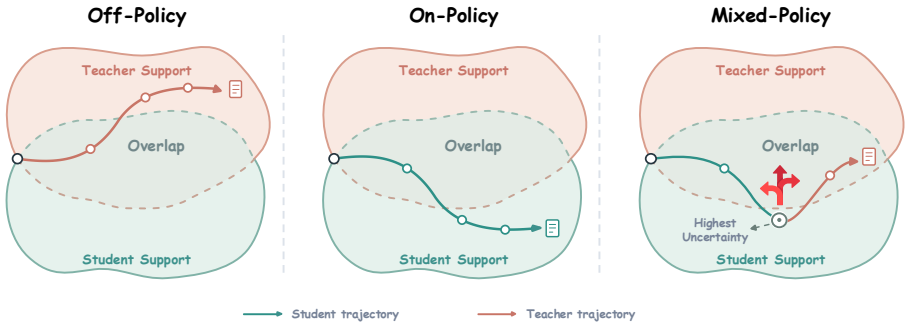


Figure 2: Conceptual comparison of trajectory construction relative to high-probability regions of the student and teacher. **Off-Policy** uses teacher-generated trajectories that may deviate from the student's typical inference distribution. **On-Policy** uses student-generated trajectories that may enter low-probability regions under the teacher. **Mixed-Policy** retains a student-generated prefix and transfers generation to the teacher at the position of highest student next-token entropy, seeking trajectories that are compatible with both policies.

Our compatibility analysis shows that, after the handoff, mixed trajectories achieve lower joint teacher–student perplexity and stronger next-token agreement than trajectories generated solely by either policy. Across mathematics, science, and tool-use tasks on four Qwen and Gemma models, MPD consistently achieves the best mean rank among OffKD, OPD, and MPD. Beyond these in-domain gains, MPD also maintains or improves upon the out-of-domain performance of OPD in the evaluated settings. Taking the tool use task as an example, MPD achieves the highest ToolAlpaca accuracy and the highest mean score on six out-of-domain benchmarks (Figure 1). The benefits of MPD also extend to privileged-information self-distillation, where the mixed-policy variant achieves a better mean rank than OPSD for every model. Our contributions are as follows:

- We propose Mixed-Policy Distillation (MPD), which uses an entropy-guided handoff strategy to combine a student-generated prefix with a teacher-generated continuation and performs full-trajectory distillation on the resulting mixed trajectory. (Sec. 3.1)

- We analyze teacher-student trajectory compatibility by comparing student-only, teacher-only, and mixed-policy trajectories. After the handoff, mixed trajectories exhibit lower joint perplexity and stronger next-token agreement than single-policy alternatives. (Sec. 3.2)
- We evaluate MPD across mathematics, science, and tool use tasks on four models from the Qwen and Gemma families, using both external teachers and privileged-information self-teachers. MPD and its self-distillation variant achieve better mean ranks than their on-policy counterparts for every model. Ablations further examine the handoff strategy and the number of handoff positions. (Sec. 4)

2 PRELIMINARIES

Off-policy knowledge distillation. Given a problem x, let π_θ denote the student policy and π_ϕ the teacher policy. We use off-policy knowledge distillation (OffKD) to denote token-level distribution matching on teacher-generated trajectories (Hinton et al., 2015; Kim & Rush, 2016; Sanh et al., 2019). Specifically, we sample $y \sim \pi_\phi(\cdot \mid x)$ and train the student to match the teacher's full next-token distribution at every position of y:

$$\mathcal{L}_{\mathrm{OffKD}}(\theta) = \mathbb{E}_{y\sim\pi_\phi(\cdot|x)}\left[\frac{1}{|y|}\sum_{t=1}^{|y|} D_{\mathrm{KL}}(\pi_\phi(\cdot \mid x, y_{<t}) \,\|\, \pi_\theta(\cdot \mid x, y_{<t}))\right], \tag{1}$$

where D_{KL} is the Kullback–Leibler divergence between next-token distributions.

On-policy distillation. On-policy distillation (OPD) samples a response $y \sim \pi_\theta(\cdot \mid x)$ from the student and then trains the student to match the teacher's next-token distributions at each token position along the student-generated response y (Agarwal et al., 2024). The OPD objective is

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{y\sim\pi_\theta(\cdot|x)}\left[\frac{1}{|y|}\sum_{t=1}^{|y|} D_{\mathrm{KL}}(\pi_\phi(\cdot \mid x, y_{<t}) \,\|\, \pi_\theta(\cdot \mid x, y_{<t}))\right], \tag{2}$$

On-policy self-distillation. On-policy self-distillation (OPSD) replaces the external teacher with the student model itself conditioned on privileged information c as additional prefixes, and the self-teacher is treated as a fixed target via the stop-gradient operator $\mathrm{sg}[\cdot]$ during training (Hübotter et al., 2026; Shenfeld et al., 2026; Zhao et al., 2026):

$$\mathcal{L}_{\mathrm{OPSD}}(\theta) = \mathbb{E}_{y\sim\pi_\theta(\cdot|x)}\left[\frac{1}{|y|}\sum_{t=1}^{|y|} D_{\mathrm{KL}}\big(\mathrm{sg}\big[\pi_{\bar\theta}(\cdot \mid x, c, y_{<t})\big] \,\|\, \pi_\theta(\cdot \mid x, y_{<t})\big)\right], \tag{3}$$

where the self-teacher parameters $\bar\theta$ are either fixed to the initial student or updated as an exponential moving average (EMA) of θ.

Throughout this paper, student rollouts are generated by the current policy π_θ. During student updates, the trajectory-generation distributions, teacher targets, and any handoff decisions (applied to our MPD method) are treated as fixed (i.e., stop gradient), while gradients are taken only through the student distributions inside the token-level KL.

The above objectives are presented using the forward KL $D_{\mathrm{KL}}(\pi_\phi \,\|\, \pi_\theta)$, while swapping its arguments gives the reverse KL $D_{\mathrm{KL}}(\pi_\theta \,\|\, \pi_\phi)$. Forward KL is mode-covering: it encourages the student to cover all tokens the teacher deems likely. Reverse KL is mode-seeking: it encourages the student to concentrate on a subset of those tokens (Gu et al., 2024; Agarwal et al., 2024). Independently of the KL direction, token-level KL can be computed over the full shared vocabulary $\mathcal{V}$ or approximated using a top-K token subset at each position when computational resources are limited.

3 MIXED-POLICY DISTILLATION

By joining a student-generated prefix with a teacher-generated continuation, MPD shapes the trajectory distribution for compatibility with both policies, rather than adopting that of either policy.

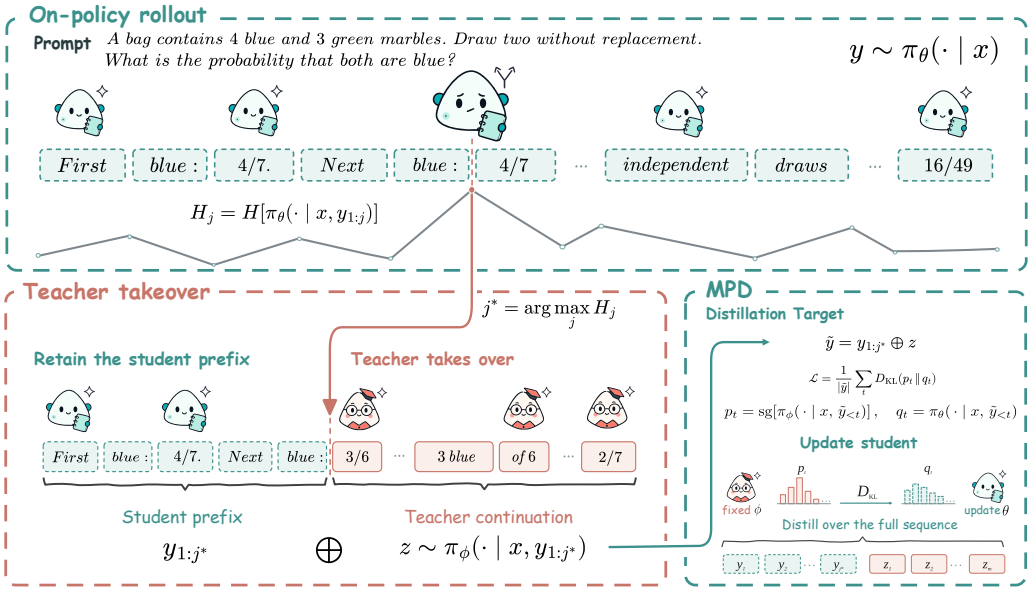


Figure 3: Overview of Mixed-Policy Distillation (MPD). **Top:** MPD selects the handoff position $j^* = \arg\max_j H_j$ with the highest student next-token entropy. **Bottom left:** The teacher generates a continuation z from the student prefix $y_{1:j^*}$. **Bottom right:** The student matches the teacher's next-token distributions over the mixed trajectory $\tilde{y} = y_{1:j^*} \oplus z$.

3.1 Mixed-Policy Trajectory Construction

MPD constructs each training response from two contiguous segments. The student first produces a rollout $y \sim \pi_\theta(\cdot \mid x)$. At the handoff position j, MPD retains $y_{1:j}$ as the student prefix. The teacher then continues from the prompt and this prefix, producing $z = (z_1, \ldots, z_m) \sim \pi_\phi(\cdot \mid x, y_{1:j})$. Each teacher token z_t is conditioned on x, the complete prefix $y_{1:j}$, and the preceding continuation $z_{<t}$. The response used for distillation is

$$\tilde{y} = y_{1:j} \oplus z = (y_1, \ldots, y_j, z_1, \ldots, z_m). \tag{4}$$

Here $\oplus$ denotes concatenation. The student controls the states in $y_{1:j}$, while the teacher controls the states created after the handoff. Section 3.3 describes how the handoff position j is selected.

Following the token-level formulation used in autoregressive distillation (Agarwal et al., 2024), MPD applies the forward-KL loss of Equation 2 to every position of the mixed response $\tilde{y}$:

$$\mathcal{L}_{\mathrm{MPD}}(\theta) = \mathbb{E}_{x \sim \mathcal{D},\, y \sim \pi_\theta(\cdot|x),\, z \sim \pi_\phi(\cdot|x, y_{1:j})} \left[\frac{1}{|\tilde{y}|} \sum_{t=1}^{|\tilde{y}|} D_{\mathrm{KL}}(\pi_\phi(\cdot \mid x, \tilde{y}_{<t}) \,\|\, \pi_\theta(\cdot \mid x, \tilde{y}_{<t})) \right], \tag{5}$$

which decomposes as $\mathcal{L}_{\mathrm{MPD}}(\theta) = \mathcal{L}_y(\theta) + \mathcal{L}_z(\theta)$, where $\mathcal{L}_y$ and $\mathcal{L}_z$ collect the terms with $t \le j$ and $t > j$, respectively. The shared denominator gives equal weight to every token in $\tilde{y}$. The first term provides teacher supervision on the student-generated prefix, and the second distills the teacher-generated continuation.

Let $v \in \mathcal{V}$ denote a candidate next token. The gradient of the token-level loss at position t can be written as (see Appendix A for the derivation)

$$\begin{aligned} &\nabla_\theta D_{\mathrm{KL}}(\pi_\phi(\cdot \mid x, \tilde{y}_{<t}) \,\|\, \pi_\theta(\cdot \mid x, \tilde{y}_{<t})) \\ &\quad = -\sum_{v \in \mathcal{V}} \pi_\theta(v \mid x, \tilde{y}_{<t}) \left(\frac{\pi_\phi(v \mid x, \tilde{y}_{<t})}{\pi_\theta(v \mid x, \tilde{y}_{<t})} - 1 \right) \nabla_\theta \log \pi_\theta(v \mid x, \tilde{y}_{<t}), \end{aligned} \tag{6}$$

which has the form of a policy gradient over the next token $v \sim \pi_\theta(\cdot \mid x, \tilde{y}_{<t})$ with advantage

$$A_t(v) = \frac{\pi_\phi(v \mid x, \tilde{y}_{<t})}{\pi_\theta(v \mid x, \tilde{y}_{<t})} - 1.$$

The advantage is positive when the teacher assigns a token higher probability than the student and negative otherwise. Figure 3 summarizes the training pipeline, covering the student rollout, the entropy-guided handoff, the teacher continuation, and distillation on the resulting mixed trajectory.

Algorithm 1 gives the complete training procedure. MPD supports three teacher providers. A PI-conditioned self-teacher is the student model conditioned on privileged information c, such as a reference solution (Hübotter et al., 2026; Shenfeld et al., 2026; Zhao et al., 2026), when generating z and computing target distributions. An improved-self teacher is the student's initial model further trained with reinforcement learning. A larger external teacher shares the student's tokenizer and vocabulary and enables transfer across model scales.

Algorithm 1 Mixed-Policy Distillation

Input: Data $\mathcal{D}$, student π_θ, teacher π_ϕ, steps T.
for $k = 1, \ldots, T$ **do**
  $\mathcal{B} \sim \mathcal{D}$, $\widehat{\mathcal{L}} \leftarrow 0$
  for all $x \in \mathcal{B}$ **do**
    // student rollout
    $y \sim \pi_\theta(\cdot \mid x)$
    // Entropy-guided handoff
    $j^* \leftarrow \arg\max_{0 \le j < |y|} H_j$
    // teacher continuation
    $z \sim \pi_\phi(\cdot \mid x, y_{1:j^*})$
    // Mixed trajectory
    $\tilde{y} \leftarrow y_{1:j^*} \oplus z$
    $\widehat{\mathcal{L}} \leftarrow \widehat{\mathcal{L}} + |\mathcal{B}|^{-1}(\mathcal{L}_y + \mathcal{L}_z)$
  end for
  $\theta \leftarrow \mathrm{OptimizerStep}(\theta, \nabla_\theta \widehat{\mathcal{L}})$
end for

3.2 Diagnostic Analysis

MPD is designed to construct distillation targets that remain compatible with both student and teacher. We test this property against student-only and teacher-only targets using two complementary signals: the likelihood of the realized target under both policies and their agreement over plausible next tokens. Both signals show that, after the handoff, mixed-policy targets are more compatible with both policies than targets from either policy alone. Appendix C provides the evaluation protocol and extends the analysis to five additional examples. It also reports separate student and teacher perplexities and compares entropy-based, KL-based, and random handoff selection.

Mixed-policy compatibility. Student-only, teacher-only, and mixed-policy targets instantiate on-policy, off-policy, and mixed-policy generation. The fixed initial student π_{θ_0} and teacher π_ϕ compute perplexity along every target, and their geometric mean measures how familiar the realized trajectory is to both policies. For a target $\tilde{y}$ and prefix length t, we compute the teacher and student perplexities of the first t target tokens and their geometric mean:

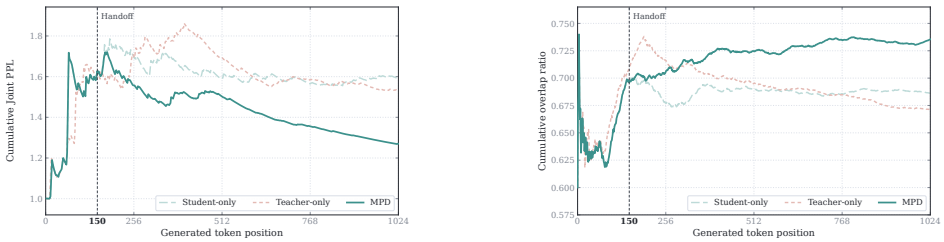
$$\mathrm{PPL}_{\psi,t} = \exp\left(-\frac{1}{t}\sum_{i=1}^{t}\log \pi_\psi(\tilde{y}_i \mid x, \tilde{y}_{<i})\right), \quad \psi \in \{\phi, \theta_0\},$$
$$\text{Joint PPL}_t = \sqrt{\mathrm{PPL}_{\phi,t}\,\mathrm{PPL}_{\theta_0,t}}. \tag{7}$$

Lower joint perplexity means that the target receives higher combined likelihood from the student and teacher. This metric evaluates the realized trajectory as a whole sequence, including the states created by its earlier tokens. In addition, we measure student–teacher consensus along the states visited by each target. At every prefix, we compute the fraction of shared top-K next-token candidates and average it over the first t positions:

$$\mathrm{Overlap}_t = \frac{1}{t}\sum_{i=1}^{t}\frac{|\,\mathrm{TopK}(\pi_{\theta_0}(\cdot \mid x, \tilde{y}_{<i})) \cap \mathrm{TopK}(\pi_\phi(\cdot \mid x, \tilde{y}_{<i}))|}{K}. \tag{8}$$

Both policies are conditioned on the same prefix $(x, \tilde{y}_{<i})$, and TopK returns the K most probable next tokens under each policy. A higher overlap means that the student and teacher agree more on which tokens are likely to be generated next along the target.

Figure 4 shows a consistent pattern in both measurements. Before the handoff, MPD shares its prefix with the student-only target, so the two curves coincide. After the handoff, the joint perplexity of MPD continues to decrease and stays below both baselines, while its top-K overlap gradually rises above them by the end of the response. The joint perplexity in Figure 4, together with the separate student and teacher perplexity curves in Appendix C, shows that mixed trajectories become more compatible with both policies as generation proceeds after the handoff: cumulative perplexities generally decrease and eventually fall below those of both single-policy baselines. Their increasing top-K overlap further indicates stronger agreement on plausible next tokens. This supports assigning the prefix and the continuation to different policies, hence our MPD scheme.



(a) Cumulative joint perplexity.

(b) Cumulative top-K overlap.

Figure 4: Trajectory compatibility of student-only (on-policy), teacher-only (off-policy), and mixed-policy targets. **Left:** Cumulative joint perplexity, the geometric mean of the student and teacher perplexities, measures how familiar each target is to both policies (lower is better). **Right:** Cumulative top-K overlap ($K = 20$) measures how much the two policies agree on plausible next tokens along each target (higher is better). Vertical dashed lines mark the MPD handoff.

3.3 Selection of Handoff Position

MPD places the handoff where the student is most uncertain about its next token. High-entropy positions tend to mark branching points in the response, where the student is hesitant about how to continue. Therefore, handing off at the student's most uncertain position keeps the confident part of the rollout as on-policy context and places teacher supervision where the student needs it.

Let $y \sim \pi_\theta(\cdot \mid x)$ be the complete student rollout. At each position j, the student's entropy over the next-token distribution following $y_{1:j}$ is

$$H_j = -\sum_{v \in \mathcal{V}} \pi_\theta(v \mid x, y_{1:j}) \log \pi_\theta(v \mid x, y_{1:j}). \tag{9}$$

Note that entropy is computed directly from the student's rollout distributions, providing a low-cost measure of next-token uncertainty. MPD selects $j^* \in \arg\max_{0 \le j < |y|} H_j$, retains $y_{1:j^*}$, and constructs the mixed response as in Equation 4. The student prefix therefore ends right before its most uncertain token, and the teacher generates the rest of the response from that point. Table 3 in Section 4.4 compares this rule with random and KL-based handoff selection.

4 Experiments

4.1 Experimental Setup

Training. We train Qwen3.5-4B, Qwen3.5-9B, Gemma3-4B, and Gemma3-12B separately on mathematics, science, and tool use, using the hard subset of OpenR1 for mathematics, the Chemistry L-3 subset of SciKnowEval for science, and ToolAlpaca for tool use, with the train–test splits of SDFT (Shenfeld et al., 2026) for the latter two. SFT is implemented with ms-swift, and all other methods are implemented with verl. All models are trained in non-thinking mode, and for each run we select the checkpoint with the best accuracy on a held-out validation set. Table 5 in Appendix B lists the training hyperparameters.

Evaluation. We evaluate on AIME25, AIME26, HMMT25, and HMMT26 for mathematics, and on the test sets of SciKnowEval-Chemistry and ToolAlpaca for science and tool use, respectively. All evaluations are conducted in non-thinking mode. We use greedy decoding for science and tool use and report accuracy. For mathematics, we estimate pass@8 from 16 sampled responses per problem with the unbiased estimator of Chen et al. (2021) (Appendix E) and report Math Avg. as the average across the four benchmarks. Mean Rank averages the ranks on Math Avg., science accuracy, and tool use accuracy within each teacher setting.

Teacher providers and comparisons. We evaluate MPD with two teacher providers. In the external-teacher setting, we adopt a DeepSeek-style teacher construction (Xu et al., 2026): each Base model is first strengthened with GRPO, and the resulting checkpoint serves as the teacher for

Table 1: Distillation with GRPO-trained teachers. Within each model and domain, distilled students share the Base initialization and a fixed GRPO-trained teacher. Base, SFT, and GRPO are listed for reference. Mathematics reports pass@8 (%); Science and Tool use report accuracy (%). Bold indicates the best result among OffKD, OPD, and MPD. Mean Rank averages their ranks across Mathematics Avg., Science, and Tool use. * marks a distillation score above the fixed teacher.

	Method		Mathematics					Science	Tool use	
			AIME 25	AIME 26	HMMT 25	HMMT 26	Avg.	SciKnow Eval	Tool Alpaca	Mean Rank ↓
Qwen3.5-4B										
Reference	Base	student init.	57.09	69.30	49.08	39.85	53.83	57.79	41.23	–
	SFT		69.64	77.84	52.32	44.35	61.04	67.46	64.94	–
	GRPO	fixed teacher	73.87	76.35	56.07	50.91	64.30	66.86	73.19	–
Distillation	OffKD	off-policy	65.37	74.59	56.48*	46.24	60.67	66.66	70.10	2.33
	OPD	on-policy	66.47	72.69	51.73	46.19	59.27	63.31	71.13	2.67
	MPD	mixed-policy	**68.09**	**74.76**	**56.77***	**46.48**	**61.53**	**68.44***	**74.22***	**1.00**
Qwen3.5-9B										
Reference	Base	student init.	67.37	75.20	59.82	50.48	63.22	50.29	45.36	–
	SFT		72.61	77.81	66.25	55.52	68.05	69.23	70.10	–
	GRPO	fixed teacher	83.03	85.71	77.67	61.80	77.05	61.73	70.10	–
Distillation	OffKD	off-policy	70.80	79.02	64.30	54.30	67.11	61.73	69.07	2.67
	OPD	on-policy	73.62	78.82	65.14	54.99	68.14	57.79	**72.16***	2.17
	MPD	mixed-policy	**74.20**	**80.75**	**67.55**	**58.91**	**70.35**	**63.31***	**72.16***	**1.17**
Gemma3-4B										
Reference	Base	student init.	20.83	21.44	1.66	6.06	12.50	21.10	41.23	–
	SFT		22.66	24.73	5.00	9.90	15.57	30.76	53.60	–
	GRPO	fixed teacher	19.62	27.37	5.00	9.05	15.26	48.91	60.82	–
Distillation	OffKD	off-policy	24.05*	24.90	**7.55***	10.26*	16.69*	48.12	61.85*	2.50
	OPD	on-policy	**24.53***	23.75	**7.55***	10.30*	16.53*	48.52	61.85*	2.50
	MPD	mixed-policy	24.20*	**25.42**	**7.55***	**10.70***	**16.97***	**49.70***	**62.88***	**1.00**
Gemma3-12B										
Reference	Base	student init.	30.55	29.21	16.44	10.60	21.70	33.72	43.29	–
	SFT		34.85	34.96	13.33	20.37	25.88	45.36	61.85	–
	GRPO	fixed teacher	34.65	37.68	19.49	13.59	26.35	62.52	70.10	–
Distillation	OffKD	off-policy	34.20	37.97*	17.51	**18.78***	27.12*	60.55	67.01	2.50
	OPD	on-policy	37.00*	35.95	**20.55***	13.11	26.65*	**61.93**	67.01	2.17
	MPD	mixed-policy	**37.88***	**39.81***	18.43	15.96*	**28.02***	60.74	**69.07**	**1.33**

OffKD (Hinton et al., 2015; Kim & Rush, 2016; Sanh et al., 2019), OPD (Agarwal et al., 2024), and MPD students initialized from the same Base. We report SFT and GRPO as standard post-training references. The primary comparison among Base, OffKD, OPD, and MPD measures the effect of distillation and mixed-policy trajectory construction under this shared teacher.

In the PI-conditioned self-teacher setting, we compare Base, OPSD (Zhao et al., 2026), and mixed-policy self-distillation (MPSD), the self-distillation instantiation of MPD. OPSD and MPSD start from the same Base model and use its PI-conditioned distribution as teacher supervision. They share the same PI, and following SDFT (Shenfeld et al., 2026), both use an exponential moving average (EMA) of the student with $\alpha = 0.01$ as the self-teacher.

4.2 Main Results

Table 1 compares the distillation methods with GRPO-trained teachers. MPD achieves the best Mean Rank on all four models. In mathematics, it attains the highest Math Avg. on every model, improving over OPD by up to 2.26 points. On tool use and science, MPD matches or exceeds OPD on all four models and three of the four models, respectively. Table 2 shows the same trend with PI-conditioned self-teachers. MPSD achieves a better Mean Rank than OPSD on all four models and improves Math Avg. by up to 3.13 points. The gains are largest on the Gemma models, reaching 12.37 points on tool use and 11.84 points on science. Appendix D further compares the training dynamics of MPD, OPD, and OffKD.

Table 2: PI-conditioned self-distillation. OPSD and MPSD share the Base initialization and privileged information. Mathematics reports pass@8 (%); Science and Tool use report accuracy (%). Bold indicates the best result between OPSD and MPSD. Mean Rank averages their ranks across Mathematics Avg., Science, and Tool use.

	Method		Mathematics					Science	Tool use	
			AIME 25	AIME 26	HMMT 25	HMMT 26	Avg.	SciKnow Eval	Tool Alpaca	Mean Rank ↓
Qwen3.5-4B										
Reference	Base	student init.	57.09	69.30	49.08	39.85	53.83	57.79	41.23	–
Distillation	OPSD	on-policy	65.73	75.54	54.99	44.09	60.09	67.85	**73.19**	1.83
Distillation	**MPSD**	mixed-policy	**68.76**	**75.86**	**56.21**	**48.15**	**62.24**	**68.84**	**73.19**	**1.17**
Qwen3.5-9B										
Reference	Base	student init.	67.37	75.20	59.82	50.48	63.22	50.29	45.36	–
Distillation	OPSD	on-policy	72.93	**83.30**	66.31	53.50	69.01	**67.06**	69.07	1.67
Distillation	**MPSD**	mixed-policy	**75.14**	80.34	**66.85**	**56.18**	**69.63**	66.66	**72.16**	**1.33**
Gemma3-4B										
Reference	Base	student init.	20.83	21.44	1.66	6.06	12.50	21.10	41.23	–
Distillation	OPSD	on-policy	22.42	20.75	3.33	**10.61**	14.28	56.01	48.45	2.00
Distillation	**MPSD**	mixed-policy	**28.50**	**24.98**	**5.88**	10.29	**17.41**	**67.85**	**56.70**	**1.00**
Gemma3-12B										
Reference	Base	student init.	30.55	29.21	16.44	10.60	21.70	33.72	43.29	–
Distillation	OPSD	on-policy	32.98	30.73	17.86	20.14	25.43	66.86	56.70	2.00
Distillation	**MPSD**	mixed-policy	**38.41**	**36.09**	**19.22**	**20.30**	**28.50**	**73.37**	**69.07**	**1.00**

4.3 In-Domain and Out-of-Domain Performance

Following SDFT (Shenfeld et al., 2026), we evaluate each checkpoint on HellaSwag, TruthfulQA, MMLU, IFEval, Winogrande, and HumanEval, and use their mean score as out-of-domain (OOD) performance. In-domain (ID) performance is the corresponding domain score from Tables 1 and 2: Math Avg., SciKnowEval accuracy, or ToolAlpaca accuracy. Figure 1 pairs ID and OOD scores after tool use training on Qwen3.5-4B and Gemma3-4B with GRPO-trained teachers. On both models, MPD achieves the highest tool use accuracy and the highest mean OOD score among the compared methods. We report the corresponding results for mathematics and science on Qwen3.5-4B and Gemma3-4B in Appendix F.

4.4 Ablations and Design Analysis

We ablate two design choices of MPD on Qwen3.5-4B in science with the GRPO-trained teacher: the handoff selection rule and the number of handoff positions. Each comparison varies only the ablated component and keeps all other settings fixed.

Handoff selection. We compare the highest-entropy position used by MPD with the position of highest teacher–student forward KL, $D_{\mathrm{KL}}(\pi_\phi \,\|\, \pi_\theta)$, and a position sampled randomly from the same candidate positions. As shown in Table 3, the highest-entropy rule achieves the best accuracy, exceeding the KL-based rule by 1.58 points and random selection by 4.15 points. This result indicates that the position at which the teacher takes over has a substantial effect on distillation quality. Moreover, the entropy rule requires only a student forward pass, whereas the KL-based rule additionally requires a teacher forward pass, making the entropy rule more effective and efficient.

Number of handoff positions. We vary $M \in \{1, 2, 4, 8\}$, the number of positions selected from each student rollout. We choose the M highest-entropy positions and independently sample one teacher continuation from each, producing M mixed trajectories per problem, whose losses are averaged. Increasing M increases the teacher continuation budget. Table 4 reports test accuracy and mean teacher-generated tokens per completed training step, with $M = 1$ as the main method. Although $M = 2$ improves accuracy by 1.18 points over $M = 1$, it roughly doubles the teacher-generated tokens per step (from 51K to 102K), and larger values of M yield no further improvement.

Table 3: Handoff selection on Qwen3.5-4B in science. Tokens are mean teacher-generated tokens per training step. †Default MPD.

Rule	Entropy†	KL	Random
Acc. (%)	**68.44**	66.86	64.29
Tokens	51,122	52,373	42,665

Table 4: Number of handoff positions M on Qwen3.5-4B in science. Tokens are mean teacher-generated tokens per training step. †Default MPD.

M	1†	2	4	8
Acc. (%)	68.44	**69.62**	68.04	68.04
Tokens	51,122	102,123	192,617	379,510

5 Related Work

Post-training for large language models. Supervised fine-tuning trains models on fixed demonstrations but does not train models on their own evolving outputs. RLHF and RLVR instead optimize model-generated trajectories using preference or verifiable outcome rewards (Ouyang et al., 2022; Shao et al., 2024; Guo et al., 2025). Distillation from a stronger teacher complements these approaches with dense token-level supervision and has been adopted in the post-training of recent foundation models (Xu et al., 2026; Yang et al., 2025; Zeng et al., 2026; Team et al., 2026).

On-policy distillation. Knowledge distillation trains the student on teacher outputs by maximizing their likelihood (Kim & Rush, 2016) or matching the teacher's token-level distributions (Hinton et al., 2015; Sanh et al., 2019). On-policy distillation instead trains on student-generated sequences: ImitKD mixes fixed and student-generated sequences (Lin et al., 2020), GKD distills directly on student samples (Agarwal et al., 2024), and MiniLLM minimizes reverse KL on student-sampled responses (Gu et al., 2024). OPSD and SDFT further replace the external teacher with the student itself conditioned on privileged information (Zhao et al., 2026; Shenfeld et al., 2026). Recent studies relate the reliability of teacher targets to the length of student-generated prefixes and to teacher–student token-support overlap (Ziheng et al., 2026; Liu et al., 2026; Li et al., 2026b), and improve OPD by selecting or truncating local teacher targets (Zhang et al., 2026a; Yang et al., 2026).

Teacher-guided distillation. Speculative KD repeatedly accepts or replaces student tokens using teacher rankings (Xu et al., 2025), whereas MPD makes a single handoff based on student entropy. Guided-OPD mixes teacher and student turns through a curriculum (Li et al., 2026a); MPD selects a handoff within a response. TRD rewrites complete student rollouts (Jiang et al., 2026), while MPD preserves the selected student prefix. Step-Level OPD generates independent teacher targets from successive student prefixes (Sun et al., 2026); MPD instead generates one continuous teacher continuation and distills over the full mixed trajectory.

Our experiments focus on whether mixed trajectories alleviate the complementary mismatches of teacher-only and student-only generation. We therefore compare primarily with OffKD and OPD under a shared teacher, student initialization, and distribution-matching objective. These endpoints provide controlled references for evaluating mixed-trajectory construction, complemented by compatibility diagnostics and handoff ablations.

6 Conclusion and Future Work

We introduced Mixed-Policy Distillation (MPD), which jointly constructs training trajectories from the student and teacher policies. MPD retains a student-generated prefix, transfers generation to the teacher at the position with the highest student next-token entropy, and distills the teacher's distributions over the complete mixed trajectory. This construction anchors each trajectory in a student-generated history while allowing the teacher to shape the continuation. Our diagnostic analysis shows that mixed trajectories achieve lower joint teacher-student perplexity and greater next-token agreement than student-only or teacher-only trajectories. Across mathematics, science, and tool use, MPD improves over its on-policy counterparts with GRPO-trained external teachers and privileged-information-conditioned self-teachers, achieving the best aggregate rank on all four models.

Our current design uses a single handoff selected by an entropy-based heuristic and requires the student and teacher to share a tokenizer and vocabulary for exact distribution matching. Future work could explore learning when and how often generation should transfer between the two policies, develop mixed-policy distillation across heterogeneous tokenizers and vocabularies, and study how teacher capacity and privileged information affect the optimal trajectory composition.

AI USE STATEMENT

In this work, we used generative AI tools to assist with drafting and editing manuscript text, improving readability, LaTeX formatting, literature search, and feedback on the presentation of the methodology and experimental analyses. The authors reviewed and verified all AI-assisted text, citations, mathematical statements, experimental claims, and formatting changes. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The method and training objective are described in Section 3, with the complete procedure summarized in Algorithm 1. Section 4.1 specifies the models, datasets, teacher providers, and evaluation protocols. Additional implementation details and the compatibility-analysis protocol are provided in Appendices B and C.

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.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374.

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

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

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, et al. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026.

Li Jiang, Haoran Xu, Yichuan Ding, and Amy Zhang. Trajectory-refined distillation. *arXiv preprint arXiv:2606.08432*, 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.

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.

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

Yanjiang Liu, Jie Lou, Xinyan Guan, Yuqiu Ji, Hongyu Lin, Ben He, Xianpei Han, Le Sun, Xing Yu, and Yaojie Lu. Your teacher can't help you here: Combating supervision fidelity decay in on-policy distillation. *arXiv preprint arXiv:2605.30833*, 2026.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

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.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. *arXiv preprint arXiv:2601.19897*, 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.

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.

Yongliang Wu, Yizhou Zhou, Ziheng Zhou, Yingzhe Peng, Xinyu Ye, Xinting Hu, Wenbo Zhu, Lu Qi, Ming-Hsuan Yang, et al. On the generalization of sft: A reinforcement learning perspective with reward rectification. In *International Conference on Learning Representations*, volume 2026, pp. 27550–27571, 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*, 2026.

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

An Yang, 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.

Zhicheng Yang, Zhijiang Guo, Yifan Song, Minrui Xu, Yongxin Wang, Yiwei Wang, Xiaodan Liang, and Jing Tang. Prune-opd: Efficient and reliable on-policy distillation for long-horizon reasoning. *arXiv preprint arXiv:2605.07804*, 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.

Dongxu Zhang, Zhichao Yang, Sepehr Janghorbani, Jun Han, Andrew Ressler II, Qian Qian, Gregory D Lyng, Sanjit Singh Batra, and Robert E Tillman. Fast and effective on-policy distillation from reasoning prefixes. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 25553–25569, 2026a.

Miaosen Zhang, Yishan Liu, Shuxia Lin, Xu Yang, Qi Dai, Chong Luo, Weihao Jiang, Peng Hou, Anxiang Zeng, Xin Geng, et al. Towards on-policy sft: Distribution discriminant theory and its applications in llm training. *arXiv preprint arXiv:2602.12222*, 2026b.

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

Zhou Ziheng, Jiaqi Li, Huacong Tang, Ying Nian Wu, and Demetri Terzopoulos. Less is more: Early stopping rollout for on-policy distillation. *arXiv preprint arXiv:2605.27028*, 2026.

APPENDIX

A GRADIENT OF THE TOKEN-LEVEL DISTILLATION LOSS

For a fixed mixed trajectory, write the full-vocabulary distributions as $p_t(v) = \pi_\phi(v \mid x, \tilde{y}_{<t})$ and $q_t(v) = \pi_\theta(v \mid x, \tilde{y}_{<t})$. Holding the teacher target and response history fixed gives

$$g_t = \nabla_\theta D_{\mathrm{KL}}(p_t \| q_t) = -\sum_{v \in \mathcal{V}} p_t(v) \nabla_\theta \log q_t(v). \tag{10}$$

Since $\sum_{v \in \mathcal{V}} q_t(v) \nabla_\theta \log q_t(v) = \nabla_\theta \sum_{v \in \mathcal{V}} q_t(v) = 0$, the same gradient can be written as

$$g_t = -\sum_{v \in \mathcal{V}} q_t(v) \left(\frac{p_t(v)}{q_t(v)} - 1 \right) \nabla_\theta \log q_t(v). \tag{11}$$

Equation 11 gives the token-level update in Equation 6. Summing it over all positions of $\tilde{y}$, with the normalization in Equation 5, gives the full MPD update. The generated tokens and handoff position are held fixed during this update.

B DISTILLATION IMPLEMENTATION

Table 5 lists the training hyperparameters for OffKD, OPD, MPD, OPSD, and MPSD. The student and teacher use the same maximum response length of 8,192 tokens. We compute the token-level forward KL exactly over the full shared vocabulary, without top-K truncation.

Each GRPO-trained teacher is obtained by training the corresponding Base model with GRPO, using the hyperparameters in Table 6. The SFT reference is trained with ms-swift on the original responses of each training set, using the hyperparameters in Table 7.

For the PI-conditioned self instantiation, the teacher additionally conditions on the reference solution c both when generating continuations and when computing distributional targets. Both policies share the architecture and tokenizer and start from the same checkpoint, $\bar{\theta}_0 = \theta_0$. Following SDFT (Shenfeld et al., 2026), the self-teacher parameters $\bar{\theta}$ track an EMA of the student: after each student update,

$$\bar{\theta} \leftarrow (1 - \alpha)\, \bar{\theta} + \alpha\, \theta, \tag{12}$$

and the updated self-teacher $\pi_{\bar{\theta}}$ supplies both continuations and distributional targets in the next step. We use one student response per problem, truncate at its highest-entropy position, and sample one teacher continuation. Each problem supplies one mixed trajectory for the update.

Entropy and distillation use the original model probabilities before the decoding transformations listed in Table 5. Teacher choice determines the continuation histories and token-level targets used for distillation. Teaching effectiveness can therefore vary across providers under the same entropy-guided handoff procedure.

C COMPATIBILITY ANALYSIS PROTOCOL

We present the first five examples in the test set.

Models and target trajectories. We conduct this analysis in the science domain (SciKnowEval-Chemistry), using the Base Qwen3.5-4B checkpoint as the student π_{θ_0} and the Qwen3.5-4B checkpoint trained with GRPO on the science training set as the teacher π_ϕ. Both models are held fixed when generating and scoring the on-policy, off-policy, and mixed-policy distillation targets. Student-only and mixed-policy targets share the retained student prefix. The same two scoring models are used for all target types, with their respective input templates.

Table 5: Training hyperparameters for OffKD, OPD, MPD, OPSD, and MPSD.

Hyperparameter	Value
Training epochs	2
Hardware	8× NVIDIA H800
Divergence	Forward KL over the full vocabulary
Optimizer	AdamW
Learning rate	1×10^{-5}
Learning-rate schedule	Cosine decay
Warmup ratio	0.1
Batch size	16 prompts
Rollouts per prompt	1
Handoffs per rollout (MPD, MPSD)	1
EMA rate α (OPSD, MPSD)	0.01
Sampling temperature	1.0
Top-p / top-k	0.95 / 20
Repetition penalty	1.05
Maximum response length (student, teacher)	8,192 tokens

Table 6: GRPO teacher hyperparameters.

Hyperparameter	Value
Training epochs	2
Optimizer	AdamW
Learning rate	5×10^{-7}
Learning-rate schedule	Constant
Batch size	16 prompts
Rollouts per prompt	8
Clip ratio	0.1
KL coefficient	0.01

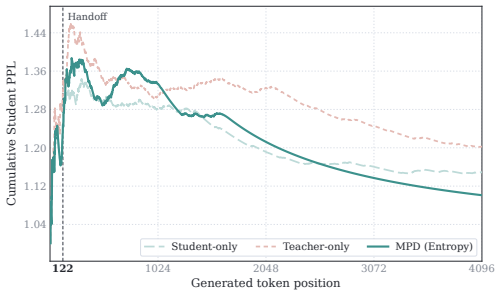
Table 7: SFT hyperparameters.

Hyperparameter	Value
Training epochs	2
Optimizer	AdamW
Learning rate	5×10^{-5}
Learning-rate schedule	Cosine decay
Warmup steps	10
Batch size per device	1
Gradient accumulation	2

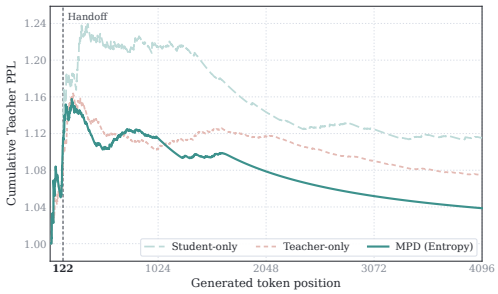
Perplexity and candidate overlap. For every target prefix, we compute student and teacher perplexities separately, then take their geometric mean as in Equation 7. The individual perplexity curves complement the aggregate comparison by showing each policy's compatibility with the targets. Top-20 overlap compares token IDs under a shared vocabulary, using probabilities before sampling truncation. We average the per-state overlap cumulatively along each target. The choice $K = 20$ matches the top-k sampling setting used in training.

Additional examples and handoff selection. Figures 5–19 compare entropy-based, KL-based, and random handoff selection for each example. Each setting is shown separately against the same student-only and teacher-only baselines. The top row reports cumulative student and teacher perplexities, and the bottom row reports cumulative joint perplexity and top-20 overlap. Vertical dashed lines mark the handoff positions.

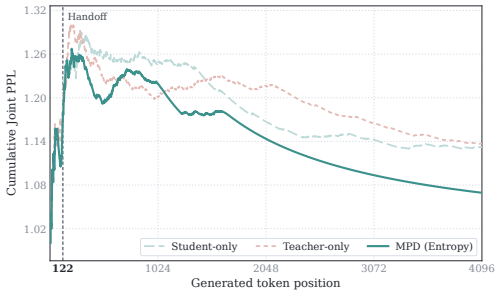
Across the five examples, the entropy-based continuation has lower cumulative joint perplexity and higher cumulative top-20 overlap than the KL-based and random continuations at the plotted endpoint. KL-based handoff selection also yields lower endpoint joint perplexity than random selection in all five examples and higher endpoint top-20 overlap in four of them. The separate student and teacher perplexity curves show how each policy contributes to the joint improvement. Overall, these compatibility trends broadly match the accuracy ordering in Table 3: entropy-based handoff selection performs best, followed by KL-based and random selection.



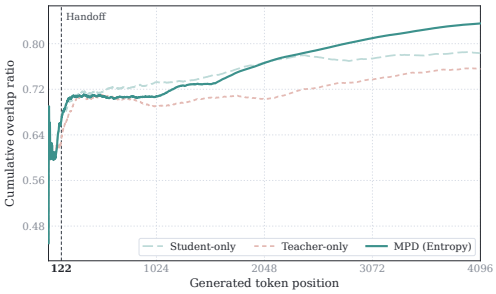
(a) Student perplexity



(b) Teacher perplexity

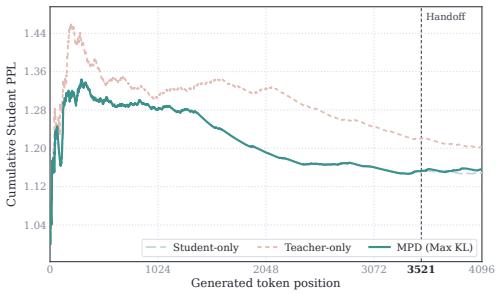


(c) Joint perplexity

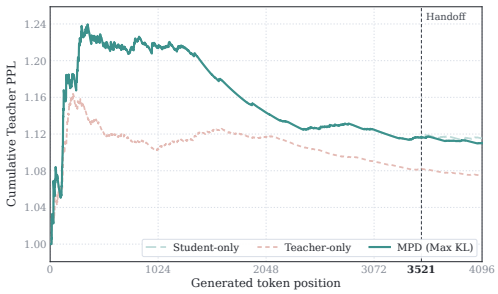


(d) Top-K overlap

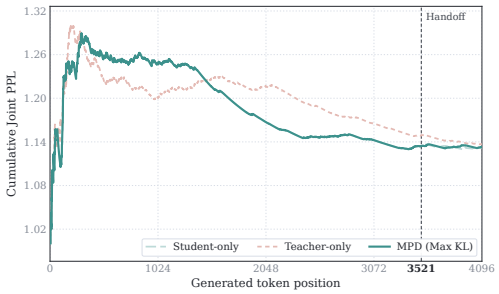
Figure 5: Sample 1 with entropy-based handoff selection. Perplexities and overlap are computed cumulatively over generated positions. Lower perplexity and higher top-K overlap ($K = 20$) are better. The vertical dashed lines mark the handoff.



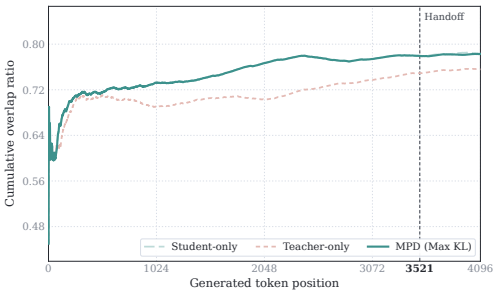
(a) Student perplexity



(b) Teacher perplexity

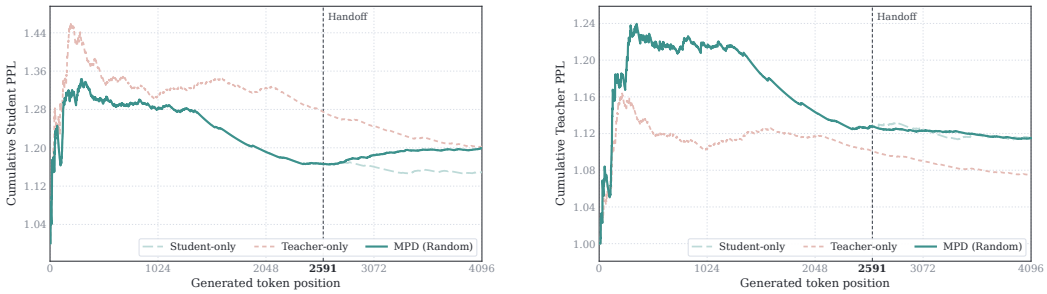


(c) Joint perplexity



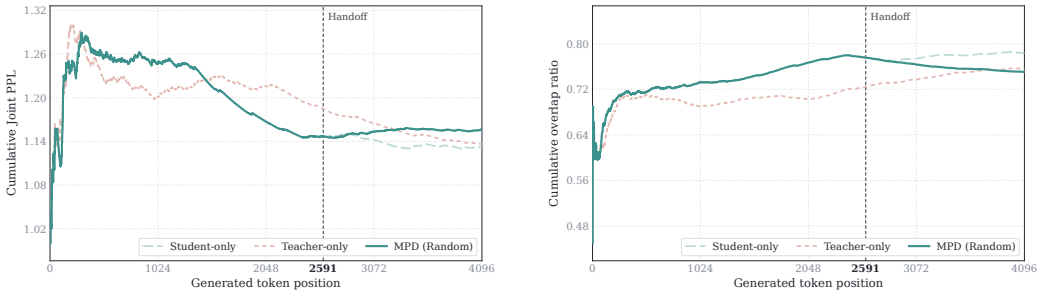
(d) Top-K overlap

Figure 6: Sample 1 with KL-based handoff selection. Perplexities and overlap are computed cumulatively over generated positions. Lower perplexity and higher top-K overlap ($K = 20$) are better. The vertical dashed lines mark the handoff.



(a) Student perplexity

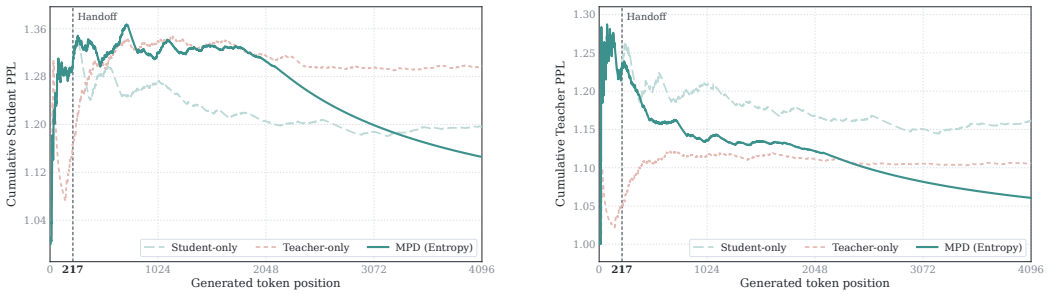
(b) Teacher perplexity



(c) Joint perplexity

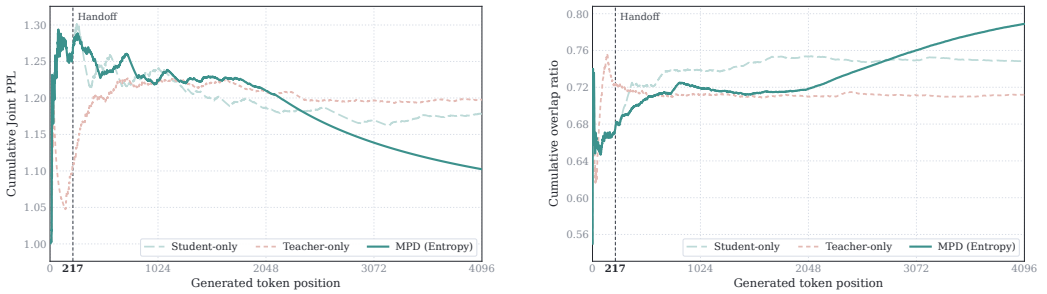
(d) Top-K overlap

Figure 7: Sample 1 with random handoff selection. Perplexities and overlap are computed cumulatively over generated positions. Lower perplexity and higher top-K overlap ($K = 20$) are better. The vertical dashed lines mark the handoff.



(a) Student perplexity

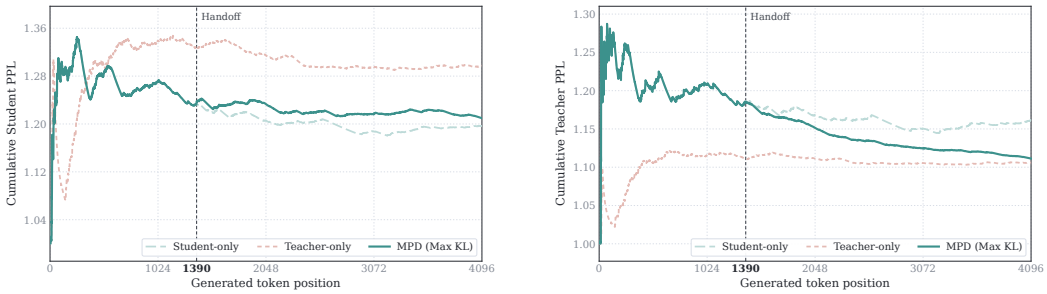
(b) Teacher perplexity



(c) Joint perplexity

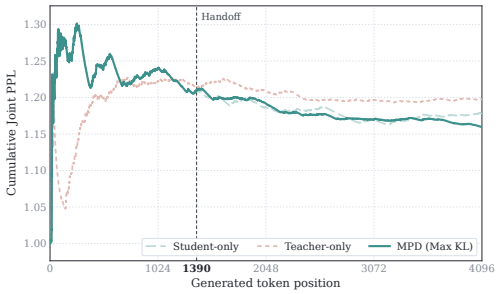
(d) Top-K overlap

Figure 8: Sample 2 with entropy-based handoff selection. Perplexities and overlap are computed cumulatively over generated positions. Lower perplexity and higher top-K overlap ($K = 20$) are better. The vertical dashed lines mark the handoff.

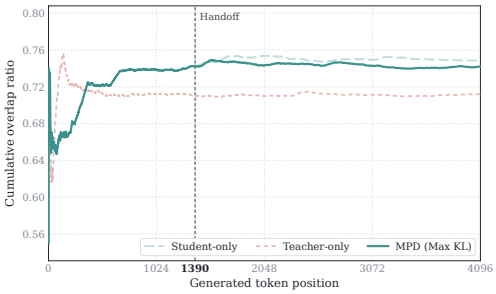


(a) Student perplexity

(b) Teacher perplexity

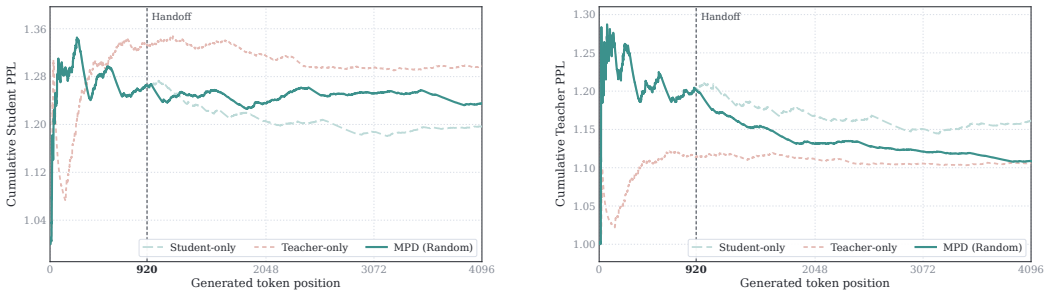


(c) Joint perplexity



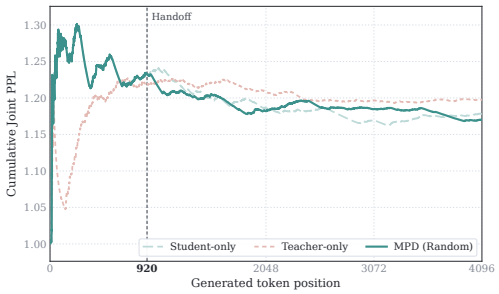
(d) Top-K overlap

Figure 9: Sample 2 with KL-based handoff selection. Perplexities and overlap are computed cumulatively over generated positions. Lower perplexity and higher top-K overlap ($K = 20$) are better. The vertical dashed lines mark the handoff.

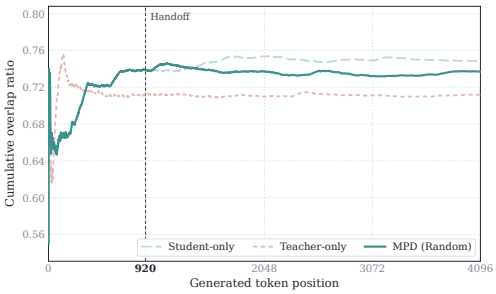


(a) Student perplexity

(b) Teacher perplexity

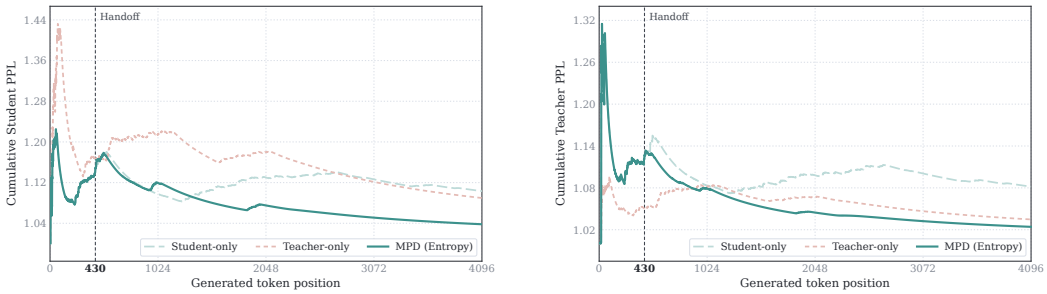


(c) Joint perplexity



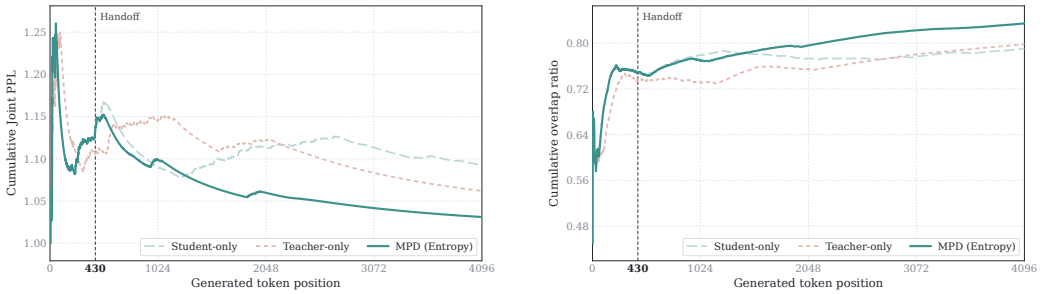
(d) Top-K overlap

Figure 10: Sample 2 with random handoff selection. Perplexities and overlap are computed cumulatively over generated positions. Lower perplexity and higher top-K overlap ($K = 20$) are better. The vertical dashed lines mark the handoff.



(a) Student perplexity

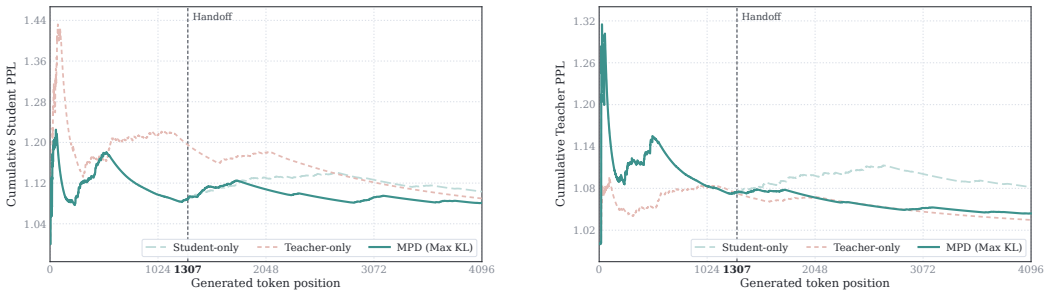
(b) Teacher perplexity



(c) Joint perplexity

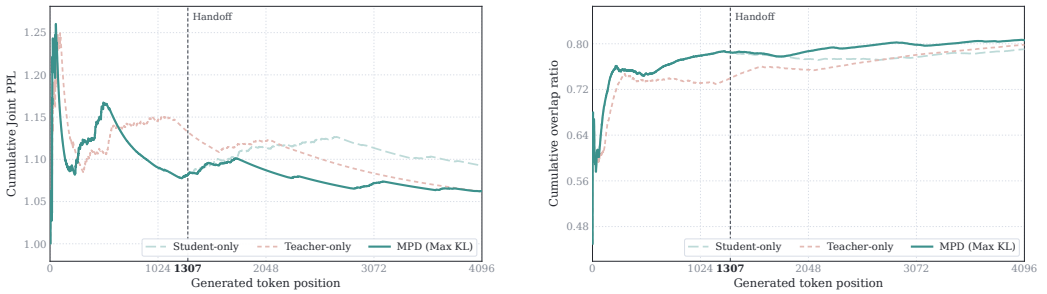
(d) Top-K overlap

Figure 11: Sample 3 with entropy-based handoff selection. Perplexities and overlap are computed cumulatively over generated positions. Lower perplexity and higher top-K overlap ($K = 20$) are better. The vertical dashed lines mark the handoff.



(a) Student perplexity

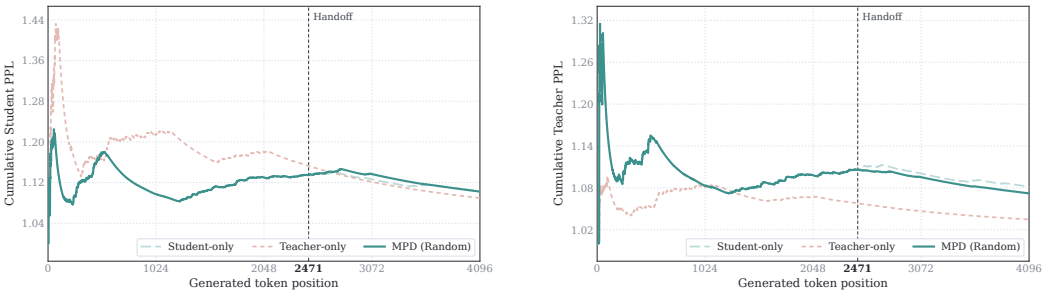
(b) Teacher perplexity



(c) Joint perplexity

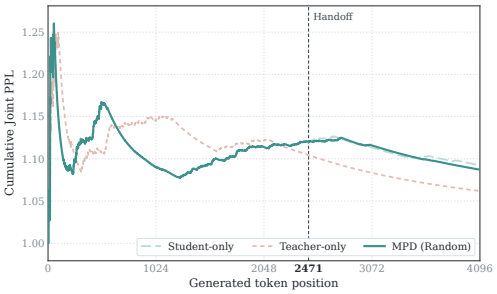
(d) Top-K overlap

Figure 12: Sample 3 with KL-based handoff selection. Perplexities and overlap are computed cumulatively over generated positions. Lower perplexity and higher top-K overlap ($K = 20$) are better. The vertical dashed lines mark the handoff.

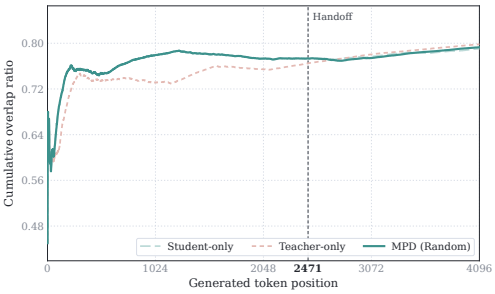


(a) Student perplexity

(b) Teacher perplexity

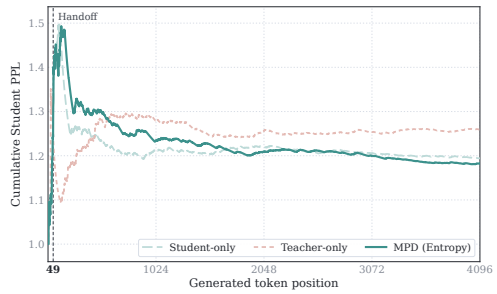


(c) Joint perplexity

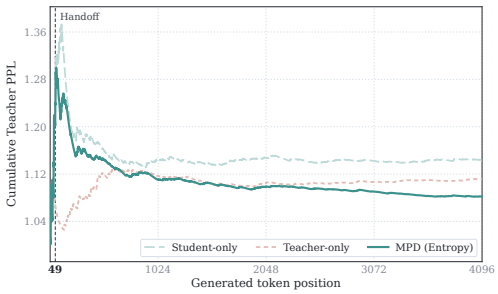


(d) Top-K overlap

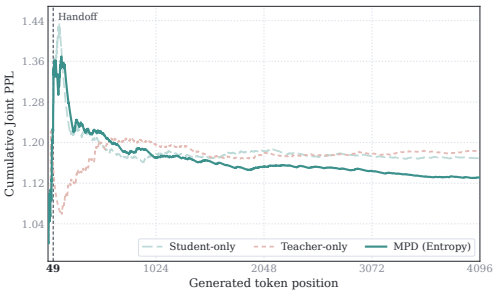
Figure 13: Sample 3 with random handoff selection. Perplexities and overlap are computed cumulatively over generated positions. Lower perplexity and higher top-K overlap ($K = 20$) are better. The vertical dashed lines mark the handoff.



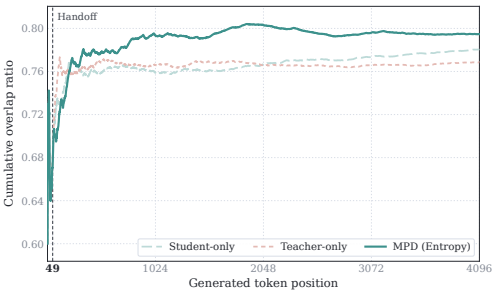
(a) Student perplexity



(b) Teacher perplexity

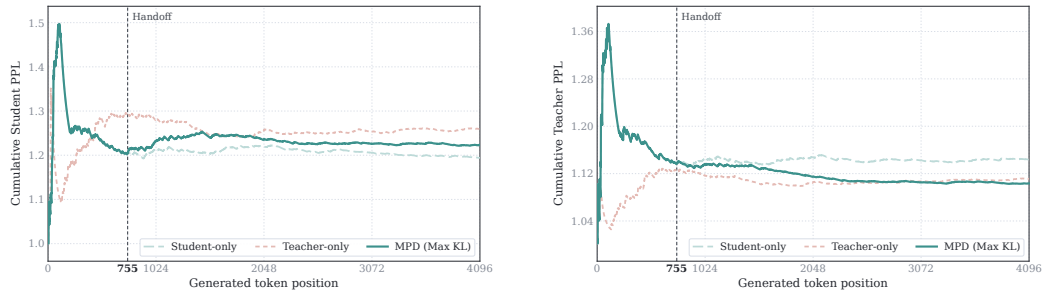


(c) Joint perplexity



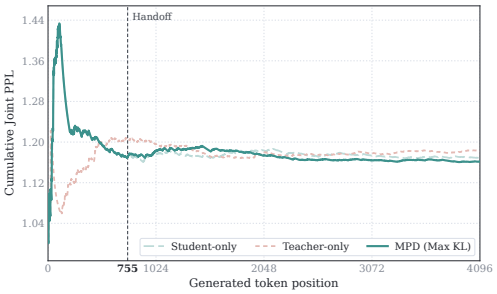
(d) Top-K overlap

Figure 14: Sample 4 with entropy-based handoff selection. Perplexities and overlap are computed cumulatively over generated positions. Lower perplexity and higher top-K overlap ($K = 20$) are better. The vertical dashed lines mark the handoff.

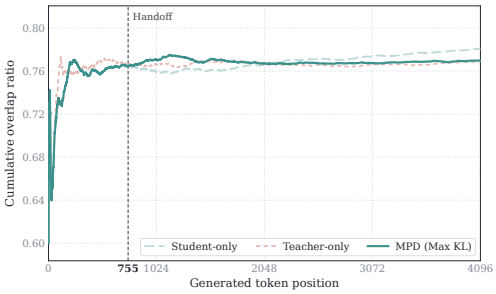


(a) Student perplexity

(b) Teacher perplexity

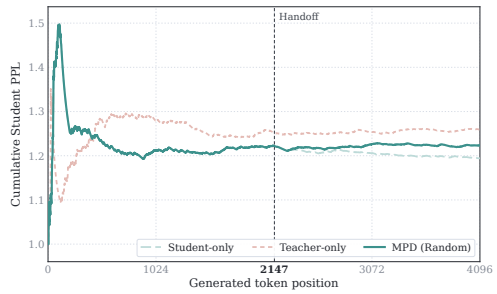


(c) Joint perplexity

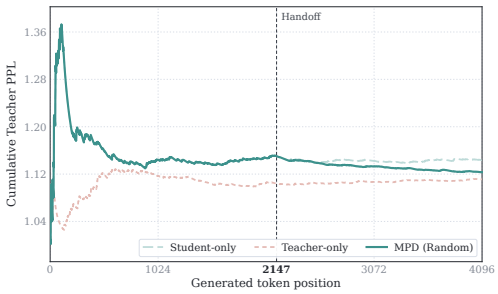


(d) Top-K overlap

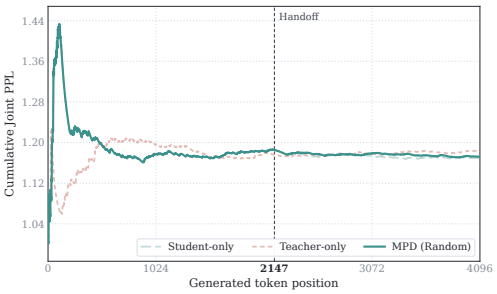
Figure 15: Sample 4 with KL-based handoff selection. Perplexities and overlap are computed cumulatively over generated positions. Lower perplexity and higher top-K overlap ($K = 20$) are better. The vertical dashed lines mark the handoff.



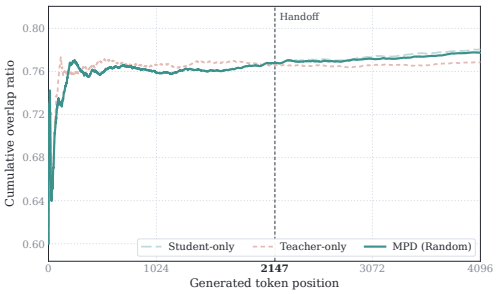
(a) Student perplexity



(b) Teacher perplexity

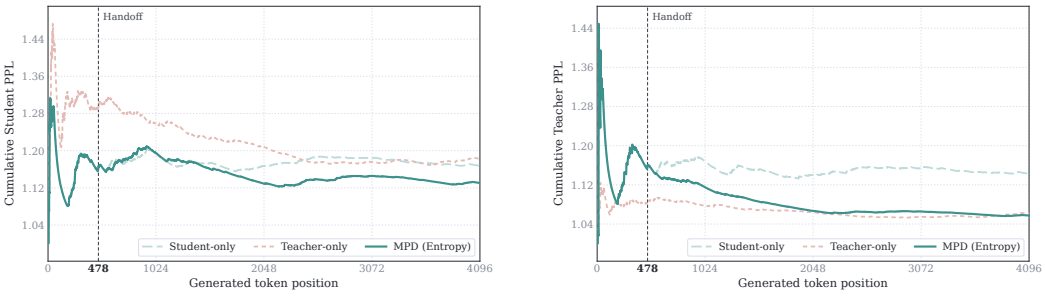


(c) Joint perplexity



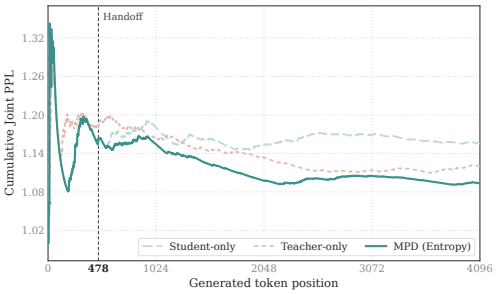
(d) Top-K overlap

Figure 16: Sample 4 with random handoff selection. Perplexities and overlap are computed cumulatively over generated positions. Lower perplexity and higher top-K overlap ($K = 20$) are better. The vertical dashed lines mark the handoff.

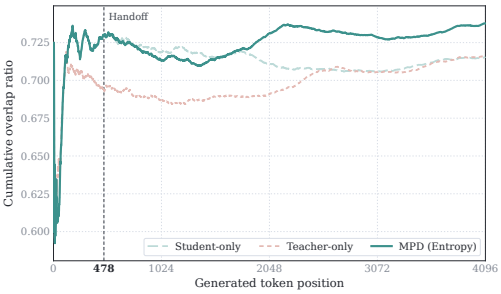


(a) Student perplexity

(b) Teacher perplexity

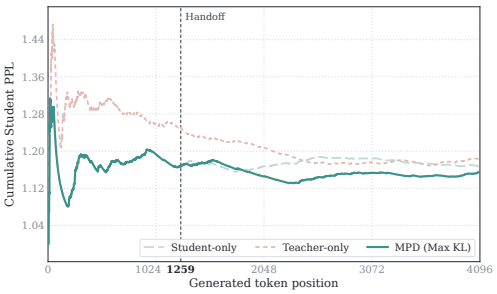


(c) Joint perplexity

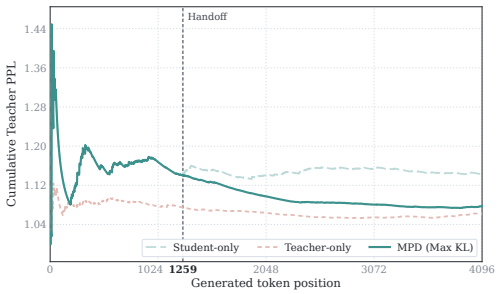


(d) Top-K overlap

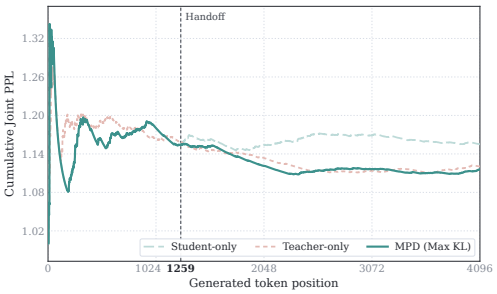
Figure 17: Sample 5 with entropy-based handoff selection. Perplexities and overlap are computed cumulatively over generated positions. Lower perplexity and higher top-K overlap ($K = 20$) are better. The vertical dashed lines mark the handoff.



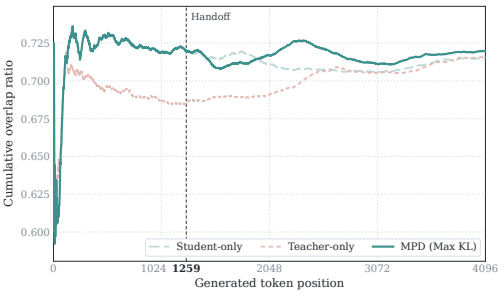
(a) Student perplexity



(b) Teacher perplexity

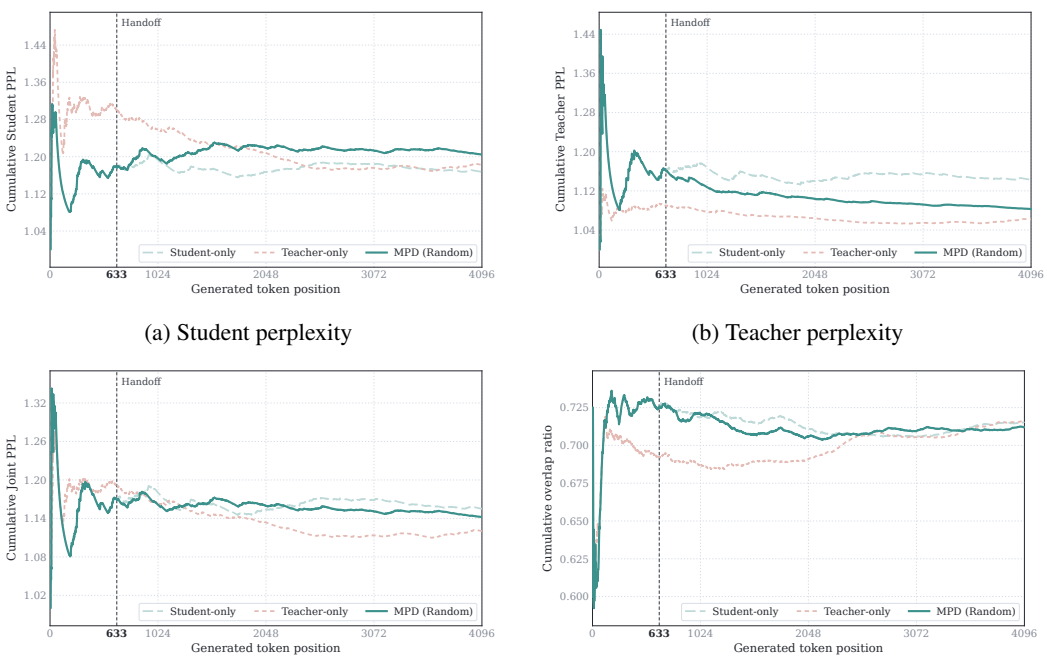


(c) Joint perplexity



(d) Top-K overlap

Figure 18: Sample 5 with KL-based handoff selection. Perplexities and overlap are computed cumulatively over generated positions. Lower perplexity and higher top-K overlap ($K = 20$) are better. The vertical dashed lines mark the handoff.



(a) Student perplexity

(b) Teacher perplexity

(c) Joint perplexity

(d) Top-K overlap

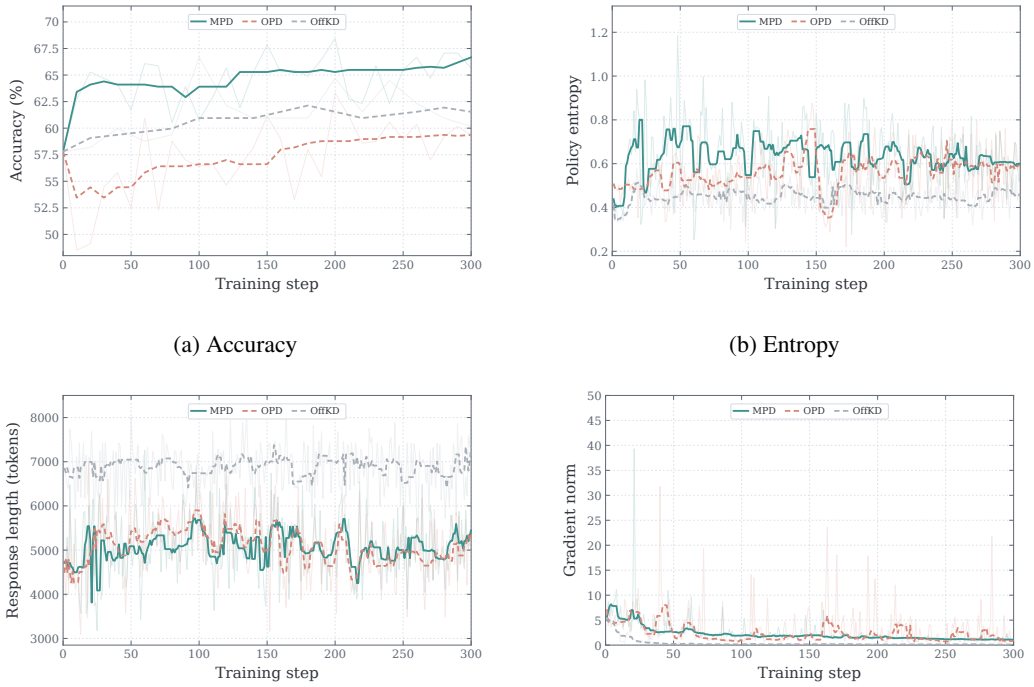
Figure 19: Sample 5 with random handoff selection. Perplexities and overlap are computed cumulatively over generated positions. Lower perplexity and higher top-K overlap ($K = 20$) are better. The vertical dashed lines mark the handoff.

D TRAINING DYNAMICS

We track MPD, OPD, and OffKD during training on Qwen3.5-4B in science. All three methods use the same GRPO-trained teacher. Figure 20 plots validation accuracy, student policy entropy, response length, and gradient norm against training steps. All methods start from the same student checkpoint and use a common evaluation schedule and decoding configuration.

At each evaluation checkpoint, we measure accuracy from the student's final answers on the validation set. We also track student policy entropy, averaged over generated response positions, mean response length, and the gradient norm of each update during training.

MPD improves faster than both baselines and stays ahead throughout training, whereas OPD first falls below the initial student, consistent with the unreliable early supervision discussed in Section 3.2. MPD also keeps the highest policy entropy and a response length similar to OPD, so its gain comes neither from a collapsed output distribution nor from longer responses. Its gradient norm is also more stable than that of OPD, which shows frequent spikes.



(a) Accuracy

(b) Entropy

(c) Response length

(d) Gradient norm

Figure 20: Training dynamics for Qwen3.5-4B on the science task. Panels show (a) validation accuracy, (b) student policy entropy, (c) mean response length, and (d) gradient norm across training steps for MPD, OPD, and OffKD. Light curves show raw measurements, and solid curves show smoothed trends.

E UNBIASED ESTIMATION OF PASS@k

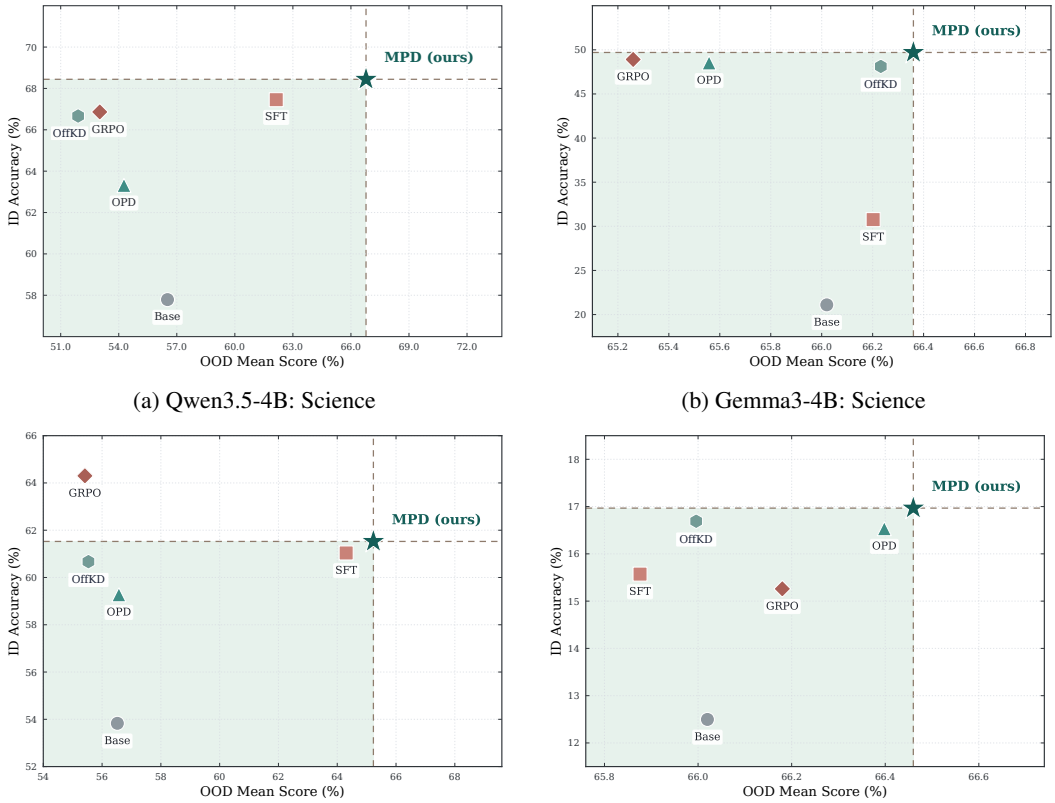
For each math problem, we sample $n = 16$ responses and count the number of correct responses c. Following Chen et al. (2021), we estimate pass@k with the unbiased estimator

$$\text{pass@}k = \mathbb{E}_{\text{problems}}\left[1 - \frac{\binom{n-c}{k}}{\binom{n}{k}}\right], \tag{13}$$

with $k = 8$. For a single problem, the bracketed term is the probability that at least one of k responses drawn without replacement from the n samples is correct. Using $n > k$ samples gives a lower-variance estimate than drawing exactly k responses.

F Additional In-Domain and Out-of-Domain Results

Figure 21 reports science and mathematics results for Qwen3.5-4B and Gemma3-4B, complementing the tool use results in Section 4.3. MPD achieves higher ID performance and mean OOD scores than both OffKD and OPD in all four settings.



(a) Qwen3.5-4B: Science

(b) Gemma3-4B: Science

(c) Qwen3.5-4B: Mathematics

(d) Gemma3-4B: Mathematics

Figure 21: ID and OOD performance after science (top) and mathematics (bottom) training on Qwen3.5-4B (left) and Gemma3-4B (right). The horizontal axis reports the mean score on the six OOD benchmarks. The vertical axis reports SciKnowEval accuracy for science and Math Avg. pass@8 for mathematics. Higher is better on both axes. Dashed lines mark MPD's scores.