ORDER-ONLY ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) trains a student on its own generations by matching the teacher's token probabilities. These probabilities entangle two kinds of information: which tokens the teacher prefers, and how confident it is. The latter reflects the teacher's own capacity and calibration, which a smaller student may be unable to reproduce. Because the teacher temperature changes these values without changing the ordering, OPD is also sensitive to the temperature choice. We ask whether the teacher's ordering of candidate tokens is enough, and propose Order-Only On-Policy Distillation (O^3PD). At each student-generated prefix, O^3PD trains the student to reproduce the order of the teacher's top-K tokens, without fitting their probability values. The objective is therefore invariant to the teacher temperature and needs no temperature tuning. Across six teacher–student pairs, O^3PD outperforms OPD on the average over HMMT25, AIME24, and AIME25 in every pair, by up to 2.9 points, with gains of up to 5.6 points on AIME25. Other ranking losses yield similar gains, so the benefit comes from ordinal supervision itself. Analysis shows that O^3PD departs further from the teacher's probabilities than OPD, yet reproduces its ranking more faithfully and yields a stronger student. The ranking, not the probability values, thus carries the knowledge that on-policy distillation needs to transfer.

1 INTRODUCTION

Knowledge distillation (KD) transfers the capabilities of a strong teacher model to a weaker student. Hinton et al. (2015) showed that the teacher's full output distribution, rather than a single hard label, is what makes this transfer effective: the probabilities assigned to competing outputs encode relationships that a hard label cannot express. This insight made distribution matching the default recipe for distillation. The student is trained to reproduce not only which output the teacher prefers, but also the exact probability that the teacher assigns to every alternative.

For autoregressive language models, the contexts in which supervision is given matter as much as the supervision itself. Sequence-level distillation trains the student with supervised fine-tuning (SFT) on teacher-generated responses (Kim & Rush, 2016). This recipe has two weaknesses. Each position supplies a single target token, so the teacher's preferences over alternatives are lost. In addition, the student is trained on teacher-generated prefixes but generates from its own prefixes at inference time, so its errors lead it into contexts it has never been trained on. On-policy distillation (OPD) removes this mismatch by querying the teacher on student-generated trajectories and training the student to match the teacher's token distribution at the states it actually visits (Gu

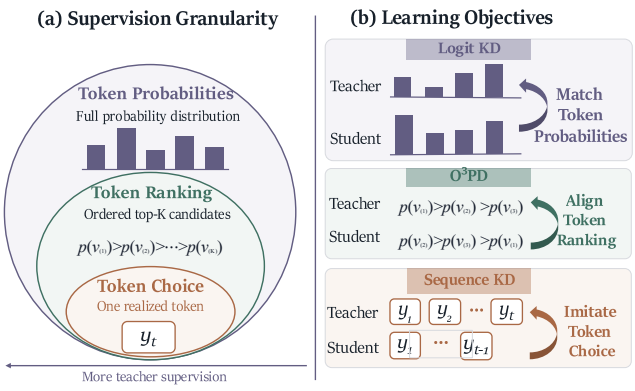

Figure 1: (a) Three nested levels of teacher supervision. (b) The corresponding objectives.

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.

et al., 2024; Agarwal et al., 2024). OPD has since become a standard approach for distilling reasoning models (Yang et al., 2025; Zhao et al., 2026).

Matching probabilities, however, asks the student for more than it needs. The teacher's probability values mix two kinds of information: which tokens it prefers, and how confident it is in that preference. The first is what the student must acquire to reason like the teacher. The second reflects the teacher's own capacity and calibration, which a much smaller student may be unable to reproduce (Cho & Hariharan, 2019; Mirzadeh et al., 2020; Fan et al., 2024). A distribution-matching objective penalizes every deviation from these values, so the student spends much of its capacity fitting a confidence profile that it cannot reach. The problem shows up clearly in practice. OPD is highly sensitive to the teacher's logit temperature, and the best temperature differs from one teacher–student pair to another (Figure 7). Moreover, OPD can degrade the student as training continues, falling below the base model it started from (Figure 3). These observations raise a basic question: *does on-policy distillation need the teacher's probabilities at all, or is the teacher's ordering of candidate tokens enough?*

We answer this question by separating three nested levels of teacher supervision: token choice, token ranking, and token probabilities (Figure 1a). A sampled target reveals a single choice. A ranked candidate list reveals which tokens the teacher considers plausible and in what order it prefers them. A probability distribution additionally fixes a numerical confidence for each token. Existing objectives occupy the two ends of this hierarchy (Figure 1b). Sequence KD imitates only the realized token, while logit KD, including distribution-matching OPD, matches the full probability distribution. Soft-target distillation shows that information about alternatives matters, but it does not show that their exact probabilities must be transferred. We argue that the middle level, namely the teacher's ranking, carries the knowledge that matters for distillation, and that the remaining probability values act as a constraint rather than a signal.

Building on this view, we propose Order-Only On-Policy Distillation (O^3PD). At each student-generated prefix, the teacher selects its top-K tokens and orders them by probability. The student is then trained with an adjacent logistic ranking loss, which requires each candidate to score higher than the next one in the teacher's order. The teacher's probabilities decide which candidates are selected and how they are ordered, but their values never enter the loss. The $K - 1$ adjacent constraints determine the complete ranking through transitivity, so O^3PD transfers the full teacher order with a minimal set of constraints while leaving the student free to set its own probability values. Because dividing logits by a positive temperature does not change their order, O^3PD is invariant to the teacher temperature by construction and needs no temperature tuning. Ordinal and relational supervision have been studied in classification distillation and in learning to rank (Huang et al., 2022; Reddi et al., 2021; Guan et al., 2025; Li & Cheng, 2026; Bassam et al., 2025), but typically alongside KL terms or confidence weights, and over fixed label sets with ground-truth labels. O^3PD uses the teacher's order as its only signal and applies it to on-policy distillation of autoregressive models, where the candidate list is rebuilt by the teacher at every student-generated prefix.

We evaluate O^3PD on six Qwen3 teacher–student configurations and three competition-level mathematics benchmarks: HMMT25, AIME24, and AIME25. Against an OPD baseline, O^3PD improves the three-benchmark average in all six configurations, by up to 2.9 points. It improves AIME25 in every configuration, by as much as 5.6 points, and it outperforms SFT and GRPO at every student scale. O^3PD also trains stably: while OPD peaks early and then declines, O^3PD keeps improving until the end of training. Our analysis explains where these gains come from. O^3PD agrees less with the teacher's probability values than OPD does, yet it reproduces the teacher's token ordering and candidate set more faithfully. It also moves the student much further from its initialization, most strongly for the smallest student, which is also the student that gains the most. Probability matching thus holds the student close to its starting point, whereas ordinal supervision lets the student reshape its distribution in the direction the teacher's preferences indicate.

Our contributions are as follows:

- We identify probability matching as an unnecessary constraint in on-policy distillation. We separate the teacher's token ordering from its probability values and show that the ordering alone provides a stronger training signal.
- We introduce O^3PD, which trains the student to reproduce the teacher's top-K token order, by default through adjacent pairwise constraints. O^3PD is invariant to the teacher

temperature, requires no temperature tuning, and remains effective across different ranking objectives.

- We show that O^{3}PD outperforms OPD on the three-benchmark average in all six teacher–student configurations and exceeds SFT and GRPO in every configuration. Through analysis, we show that O^{3}PD learns the teacher's ranking rather than its probabilities, and that it allows the smallest student to move substantially further from its initialization.

2 Related Work

Knowledge distillation (KD) can be characterized along two largely orthogonal axes: *where* supervision is applied (the prefix states on which the student is trained) and *what* teacher information is transferred at each state.

2.1 LLM Distillation: From Teacher States to Student States

Early LLM distillation trains the student on fixed sequences. Sequence-level distillation uses teacher-generated responses as targets (Kim & Rush, 2016), optionally enriched with rationales (Hsieh et al., 2023) or organized as preferences (Zhang et al., 2025), while white-box variants match the teacher's token distributions along such sequences (Hinton et al., 2015; Kim & Rush, 2016). In both cases, the student is supervised on teacher- or data-induced prefixes, whereas at inference it conditions on its own generations, creating a train–inference mismatch.

On-policy methods mitigate this mismatch by querying the teacher on student-generated states, minimizing reverse KL (Gu et al., 2024), generalized divergences (Agarwal et al., 2024), or skewed divergences with adaptive replay (Ko et al., 2024). Training on student states, however, introduces a new difficulty: when student trajectories are of low quality, teacher supervision on them becomes less reliable. SWITCH therefore lets the teacher intervene in generation when teacher–student discrepancies grow (Koo et al., 2025), and SKD replaces student tokens that fall outside the teacher's top-K (Xu et al., 2025); in SKD, the teacher's ranking decides which tokens enter the trajectory, but the training objective remains distribution matching. Complementary work improves efficiency through difficulty-aware sample selection (He et al., 2025) and curricula (Zhang & Liu, 2025). Across these methods, what changes is the states, the divergence, the trajectory construction, or the data schedule; the per-state target remains the teacher's probability distribution. O^{3}PD keeps on-policy states but replaces this target with the teacher's top-K ranking.

2.2 Distillation Beyond Probability Matching

Matching the teacher's distribution requires the student to reproduce not only which candidates the teacher prefers but also how confident the teacher is. Relaxations of this target can be characterized by the transformations of teacher scores that leave the target unchanged. KL is invariant only to logit shifts. Logit Standardization (Sun et al., 2024) and DIST (Huang et al., 2022) additionally remove scale and are invariant to affine transformations, but still consume continuous scores: a nonlinear yet order-preserving re-scoring of the teacher changes their targets. A purely ordinal target is invariant to all strictly monotone transformations and transfers the teacher's preferences without its confidence.

Ranking-based methods move toward such ordinal targets but differ from O^{3}PD in either the signal or the setting. On the signal side, RKKD adds a Kendall's τ loss as an auxiliary to KL, so teacher probabilities still enter the objective (Guan et al., 2025), and PLD, a Plackett–Luce loss akin to ListMLE (Xia et al., 2008), weights each ranked choice by teacher confidence (Bassam et al., 2025). On the setting side, these methods, as well as OPKD's pairwise ordering objectives (Li & Cheng, 2026), are designed for classification over a fixed label set with ground-truth labels, and RankDistil and DGR for ranking fixed pools of items or passages (Reddi et al., 2021; Li et al., 2024); at student-generated decoding states, by contrast, the candidate set changes from state to state and no ground-truth token exists. Conversely, BiLD operates in LLM distillation and restricts supervision to the teacher's top-K logits, but matches their pairwise differences, which are shift- but not scale-invariant and thus still encode preference strength (Li et al., 2025). O^{3}PD satisfies both requirements: at each student-generated state, it keeps only the identities and order of the teacher's top-K tokens.

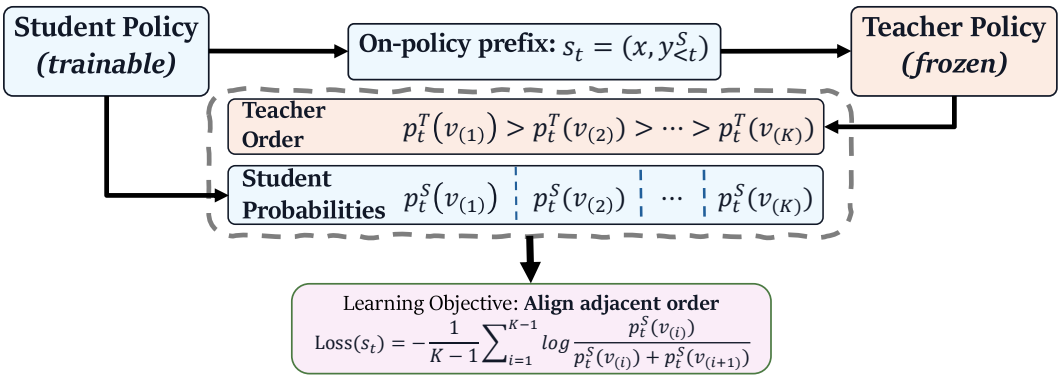

Figure 2: Overview of O^{3}PD. The figure shows the default adjacent logistic loss, any other order-only objective can take its place (Section 3.3).

3 Order-Only On-Policy Distillation

Knowledge distillation can transfer different forms of teacher supervision, from generated sequences to token-level probability distributions. We study an intermediate choice: transferring the teacher's candidate tokens and their relative ordering without fitting the teacher's probability magnitudes. We instantiate this choice as Order-Only On-Policy Distillation (O^{3}PD), which learns from teacher rankings on student-generated contexts.

3.1 From Sequence and Distribution Targets to Ordinal Supervision

Sequence-level knowledge distillation (Sequence KD) transfers knowledge through teacher-generated responses. Given a prompt x and a teacher response y^T, the student minimizes

$$\mathcal{L}_{\text{seq}}(x, y^T) = -\frac{1}{|y^T|}\sum_{t=1}^{|y^T|} \log \pi_\theta(y_t^T \mid x, y_{<t}^T). \tag{1}$$

Table 1: Comparison of KD methods.

Method	On-policy	Temperature invariant	Magnitude-free
Sequence KD	✗	✗	✓
OPD	✓	✗	✗
O^{3}PD	✓	✓	✓

This objective trains the student to reproduce the teacher's generations, but represents supervision as a selected token at each prefix without explicitly retaining preferences among alternative candidates. Moreover, training follows teacher-generated prefixes rather than output from the current student.

On-policy distillation (OPD) instead obtains teacher supervision on student-generated contexts. The student samples responses from its current policy, and the teacher provides predictive distributions at the resulting prefixes. A representative distribution-matching objective is

$$\mathcal{L}_{\text{OPD}}(s) = \text{JSD}\big(p_K^T(\cdot \mid s) \,\|\, \pi_{\theta,K}(\cdot \mid s)\big), \tag{2}$$

This supervision preserves the teacher's judgments over multiple candidates, but couples their relative ordering with the strength of their assigned probabilities. Even when candidate rankings remain unchanged, different probability assignments yield different training signals. Changing the teacher temperature likewise changes the target distribution while preserving its ordering, introducing an additional temperature-selection problem. When teacher and student differ substantially in reasoning behavior or capability, the teacher's probability assignments may also be difficult for the student to reproduce, complicating knowledge transfer. This motivates a question: can we transfer the teacher's preferences without requiring matching their probability magnitudes?

O^{3}PD separates these two forms of information. At each student-generated prefix, we extract the teacher's top-K candidates and their ordering, then train the student through adjacent comparisons. The candidate set determines which tokens are compared, and the ordering determines the preferred direction; teacher probability magnitudes are not fitting targets. This design has three properties.

First, it retains *on-policy training*, providing supervision on contexts encountered by the current student. Second, it yields *teacher-temperature invariant supervision*: at a fixed context, positive temperature scaling preserves the candidate set and its ordering, leaving the O^{3}PD objective unchanged. This removes the need to select a teacher temperature. More generally, because the objective depends on the teacher only through its top-K ordering, it is invariant to any strictly increasing transformation of the teacher's scores at each context, of which temperature scaling is a special case. Finally, it provides *magnitude-free supervision*: the objective specifies which candidate should be preferred but prescribes no target for how strongly, so the student is never required to reproduce the teacher's probability gaps. Table 1 summarizes the comparison.

3.2 Learning through Adjacent Comparisons

For a sampled response y^S, let $s_t = (x, y^S_{<t})$ denote the context at position t, z^S_t the student logits with $p^S_t = \mathrm{softmax}(z^S_t)$, and $p^T_t = p^T(\cdot \mid s_t)$ the teacher distribution. We write $(v_{t,(1)}, \ldots, v_{t,(K)}) = \mathrm{TopKOrder}(p^T_t)$ for the teacher's top-K candidates, in decreasing order of teacher preference.

O^{3}PD supervises the student through preferences between adjacent teacher candidates. For the candidates ranked i and $i{+}1$, the student's preference is the probability of selecting the higher-ranked one when the choice is restricted to the pair:

$$r^S_{t,i} = \frac{p^S_t(v_{t,(i)})}{p^S_t(v_{t,(i)}) + p^S_t(v_{t,(i+1)})} = \sigma\big(\Delta^S_{t,i}\big), \qquad \Delta^S_{t,i} = z^S_t(v_{t,(i)}) - z^S_t(v_{t,(i+1)}). \tag{3}$$

Because the softmax normalizer cancels, this Bradley–Terry preference depends only on the logit difference, and we minimize $\ell^{\mathrm{adj}}_{t,i} = -\log \sigma(\Delta^S_{t,i})$. Adjacent comparisons lose no ordinal information: if all $K-1$ adjacent inequalities hold, transitivity implies the full teacher ranking, so they form a minimal description of the order, compared with the $K(K-1)/2$ relations of all-pairs objectives. We average over positions and adjacent pairs within each response, and then over the minibatch:

$$\ell(x) = \frac{1}{|y^S|(K-1)} \sum_{t=1}^{|y^S|} \sum_{i=1}^{K-1} \ell^{\mathrm{adj}}_{t,i}, \qquad \mathcal{L}_{\mathrm{adj}} = \frac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} \ell(x). \tag{4}$$

Sampled responses and teacher rankings are held fixed during optimization (Algorithm 1).

The gradient $\partial \ell^{\mathrm{adj}}_{t,i} / \partial \Delta^S_{t,i} = -\sigma(-\Delta^S_{t,i})$ is largest for reversed pairs and decays toward zero once the preferred order is established. The loss still favors larger margins, but prescribes no target for them. Distribution matching, by contrast, has logit gradient $p^S - p^T$, which vanishes only when the student reproduces the teacher's probabilities exactly.

3.3 Order-Only as a Distillation Principle

Table 2: Order-only objectives. $\mathcal{A}$: adjacent pairs; $\mathcal{P}$: all ordered pairs; $w_{ij} = 1/i$.

Objective	Comparisons	Structure	Loss
Adjacent Logistic (O^{3}PD)	$\mathcal{A}$, $K-1$ pairs	Pairwise	$\mathrm{Avg}_{\mathcal{A}}[-\log \sigma(\Delta_{ij})]$
Logistic	$\mathcal{P}$, $K(K-1)/2$ pairs	Pairwise	$\mathrm{Avg}_{\mathcal{P}}[-\log \sigma(\Delta_{ij})]$
Top-Weighted Logistic	$\mathcal{P}$, weighted	Pairwise	$\mathrm{Avg}_{\mathcal{P},w}[-\log \sigma(\Delta_{ij})]$
ListMLE	Full list	Listwise	$-\frac{1}{K-1}\sum_{i=1}^{K-1} \log \frac{e^{z_i}}{\sum_{j=i}^{K} e^{z_j}}$

The adjacent Bradley–Terry loss is only one way to impose the teacher's order. Any objective that accesses the teacher only through its top-K candidates and their ordering, and prescribes no target for the student's score gaps, uses the same supervision as O^{3}PD and inherits its three properties from Section 3.1: on-policy training, teacher-temperature invariance, and magnitude-free supervision. Order-only is thus a principle about what the teacher provides, and O^{3}PD is one instance of it. We examine three further instances that vary which comparisons are supervised and whether the order is imposed pair by pair or as a whole (Table 2). Dropping the position index, we write $z_i = z^S(v_{(i)})$

and $\Delta_{ij} = z_i - z_j$, so that $\Delta^S_{t,i} = \Delta_{i,i+1}$; $\mathcal{P} = \{(i,j) : i < j\}$ and $\mathcal{A} = \{(i, i+1)\} \subset \mathcal{P}$, and $\mathrm{Avg}_{\mathcal{S},w}$ denotes a w-weighted average over $\mathcal{S}$ (unweighted when w is omitted).

Logistic and Top-Weighted Logistic keep the Bradley–Terry loss (Bradley & Terry, 1952) and change only the comparison set. Logistic supervises all $K(K-1)/2$ pairs, directly penalizing coarse violations such as the first candidate against the last, which the adjacent loss constrains only through transitivity. Top-Weighted Logistic weights each pair by the reciprocal rank of its preferred candidate, concentrating supervision on the head of the list, which matters most for decoding. ListMLE instead imposes the order as a whole: under the Plackett–Luce model (Plackett, 1975; Luce, 2005), it treats the ranking as a sequence of choices, each selecting the next candidate from those not yet ranked. Since Plackett–Luce reduces to Bradley–Terry on any two candidates, ListMLE expresses the same preferences as O^{3}PD, but over the full list at once. The choice among these objectives therefore determines how the order is imposed, not what is transferred.

4 Experiments

4.1 Experimental Setup

Models and data. We distill within the Qwen3 family (Yang et al., 2025). Each student (1.7B, 4B, or 8B) is paired with every larger teacher (4B, 8B, or 14B), giving six teacher–student configurations. The teacher is kept frozen, and all methods are trained on the mathematical reasoning subset of OpenThoughts (Guha et al., 2026). Full training details are given in Appendix A.

Evaluation. We evaluate on three competition-level benchmarks, HMMT25, AIME24, and AIME25, and report Avg@12: the accuracy of the final answer averaged over 12 sampled responses per problem, drawn as four samples under each of three seeds. Average denotes the unweighted mean over the three benchmarks, and all differences are reported in absolute points. For OPD and O^{3}PD, we evaluate every 20 training steps and report the checkpoint with the highest Average. This selection can only favor a method whose performance declines during training. For the 8B→1.7B pair, OPD peaks early and then degrades while O^{3}PD keeps improving (Figure 3), so comparing final checkpoints instead would widen the gap in favor of O^{3}PD.

Baselines. We compare O^{3}PD with the base student, SFT, GRPO (Shao et al., 2024), and OPD (Gu et al., 2024; Agarwal et al., 2024). SFT and GRPO results are taken from Zhao et al. (2026), which uses the same student models, training data, and evaluation protocol. OPD is our primary baseline and is trained under exactly the same configurations as O^{3}PD. It minimizes the Jensen–Shannon divergence between the student and teacher distributions restricted to the teacher's top-K tokens, with $K = 8$ and teacher temperature $T = 1$, so that both methods see the same candidate set at every prefix and differ only in whether the probability values enter the loss.

4.2 Main Results

O^{3}PD consistently outperforms baselines. Table 3 shows that O^{3}PD achieves a higher average than OPD in all six configurations. The gain is modest for the 8B student (+0.6) and larger for the smaller students: +2.3 and +1.9 for the 4B student (14B and 8B teachers), and +2.6 and +2.9 for the 1.7B student with the 8B and 4B teachers. The exception is the 14B→1.7B pair, which improves by only 0.3, suggesting that when the capacity gap is largest, even the teacher's ranking becomes harder for the student to reproduce. The gains concentrate on AIME. O^{3}PD improves AIME25 in every configuration, by 3.4 points on average and up to 5.6. It also gives the largest single gain in the table, +6.4 on AIME24 for 4B→1.7B. Overall, O^{3}PD improves over OPD in 14 of the 18 comparisons. O^{3}PD also exceeds the base model, SFT, and GRPO in average score in every configuration. For the 1.7B student, for example, it reaches 40.9, compared with 37.7 for GRPO and 35.8 for SFT. SFT falls below the base model at every scale, which highlights the value of on-policy supervision.

O^{3}PD improves steadily throughout training. Figure 3 tracks both methods for the 8B→1.7B pair. OPD peaks at step 20 (38.1) and then declines below the base student (37.7). O^{3}PD instead keeps improving and reaches its best mean (40.7) at the final step. The contrast is sharpest on AIME25, where O^{3}PD climbs from 35.6 to 43.1 while OPD falls back to its starting point. The

Table 3: Main results on mathematical reasoning (Avg@12, %). Values in parentheses are differences relative to OPD with the same teacher–student pair. Gray rows are non-distillation references; SFT and GRPO results are taken from Zhao et al. (2026).

Student	Teacher	Method	HMMT25	AIME24	AIME25	Average
Qwen3-8B	–	Base	44.4	75.8	66.4	62.2
	–	SFT	42.9	72.3	64.2	59.8
	–	GRPO	46.7	76.4	68.9	64.0
	Qwen3-14B	OPD	46.1	75.8	70.0	64.0
		O^{3}PD	46.7 (+0.6)	76.7 (+0.9)	70.3 (+0.3)	64.6 (+0.6)
Qwen3-4B	–	Base	43.9	75.8	64.7	61.5
	–	SFT	43.4	70.2	62.3	58.6
	–	GRPO	44.4	75.6	68.1	62.7
	Qwen3-14B	OPD	42.2	74.4	66.1	60.9
		O^{3}PD	45.0 (+2.8)	74.2 (−0.2)	70.3 (+4.2)	63.2 (+2.3)
	Qwen3-8B	OPD	43.1	73.9	66.9	61.3
		O^{3}PD	44.2 (+1.1)	74.7 (+0.8)	70.6 (+3.7)	63.2 (+1.9)
Qwen3-1.7B	–	Base	24.4	53.1	35.6	37.7
	–	SFT	22.7	48.4	36.3	35.8
	–	GRPO	23.7	51.1	38.3	37.7
	Qwen3-14B	OPD	25.6	52.2	39.4	39.1
		O^{3}PD	26.1 (+0.5)	51.4 (−0.8)	40.8 (+1.4)	39.4 (+0.3)
	Qwen3-8B	OPD	27.2	49.4	37.8	38.1
		O^{3}PD	27.2 (0.0)	51.9 (+2.5)	43.1 (+5.3)	40.7 (+2.6)
	Qwen3-4B	OPD	27.8	48.9	37.2	38.0
		O^{3}PD	24.7 (−3.1)	55.3 (+6.4)	42.8 (+5.6)	40.9 (+2.9)

gains of O^{3}PD therefore do not depend on checkpoint selection: at the end of training, the gap between the two methods exceeds the 2.6-point gap between their best checkpoints.

Smaller teachers suffice. O^{3}PD with a smaller teacher outperforms OPD with the 14B teacher, both for the 1.7B student (40.9 with the 4B teacher vs. 39.1) and for the 4B student (63.2 with the 8B teacher vs. 60.9). O^{3}PD thus achieves stronger distillation with a cheaper teacher.

O^{3}PD removes temperature tuning. For OPD, the teacher temperature is a consequential hyperparameter. Appendix Figure 7 sweeps $T \in \{0.25, 0.5, 1, 2, 4, 8\}$ for every teacher–student pair, together with an additional Qwen3-0.6B student. No single value is best across pairs: the optimum ranges from $T = 0.25$ for 8B→1.7B to $T = 2$ for 14B→8B and 4B→1.7B. Even after only three training steps, the choice of T alone moves HMMT25 accuracy by up to 6.4 points for students of 1.7B and above (8B→4B: 44.7 at $T = 1$ vs. 38.3 at $T = 8$). Sensitivity grows as the student shrinks, and the 0.6B student collapses at high temperatures: 14B→0.6B falls from 11.9 at $T = 1$ to 1.1 at $T = 8$, and 1.7B→0.6B drops to 0.3. Because dividing the teacher logits by any $T > 0$ leaves the top-K ordering unchanged, O^{3}PD receives identical targets at every temperature.

4.3 Ablation Study

We ablate the ranking objective and the candidate-set size K on the 8B→1.7B pair, training each variant for 100 steps (Table 4); further details are in Appendix A.

Ranking objective. Besides Adjacent Logistic, we evaluate the Logistic, Top-Weighted Logistic, and ListMLE variants of O^{3}PD. All variants outperform OPD at $K = 8$, by 1.3 to 2.6 points, so ordinal supervision is robust to the choice of ranking loss. Adjacent Logistic and ListMLE perform best, within 0.1 points of each other, even though Adjacent Logistic supervises only 7 of the 28 pairs used by Logistic. By transitivity, the adjacent relations already fix the full ranking, so supervising the remaining pairs brings no further gain. We use Adjacent Logistic by default, as it matches the best result with the fewest constraints.

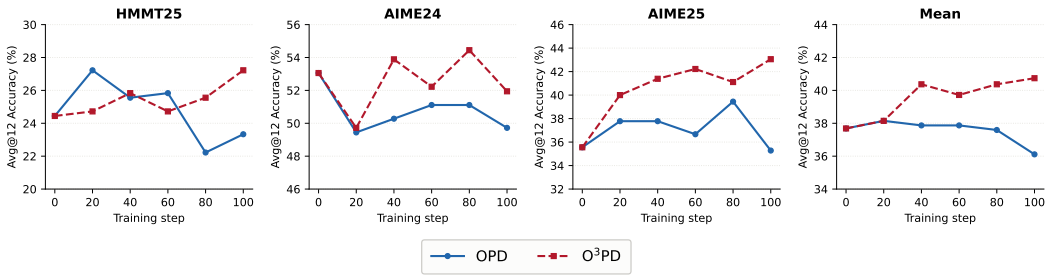

Figure 3: Performance curve of OPD and O^{3}PD for the 8B→1.7B pair. Avg@12 is evaluated every 20 training steps; step 0 is the base student.

Table 4: Ablations on the 8B→1.7B pair (100 steps). Adjacent with K=8 is the default O^{3}PD setting; Top-Weighted denotes Top-Weighted Logistic.

Method	HMMT25	AIME24	AIME25	Avg.
OPD (K=8)	27.2	49.4	37.8	38.1
Adjacent, varying K				
K=4	26.1	53.1	40.8	40.0
K=8	27.2	51.9	**43.1**	**40.7**
K=16	24.7	51.9	42.5	39.7
Other objectives, K=8				
Logistic	25.6	53.3	41.7	40.2
Top-Weighted	24.2	53.3	40.6	39.4
ListMLE	26.4	**53.6**	41.7	40.6

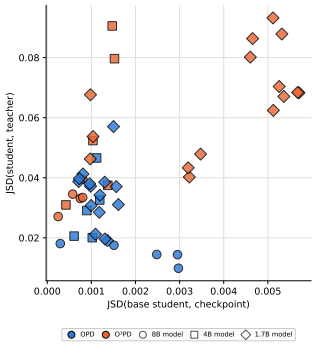

Figure 4: Training checkpoints: JSD to the base student vs. JSD to the teacher.

Candidate-set size. All three sizes outperform OPD, by 1.6 to 2.6 points, and $K = 8$ performs best. Neither a shorter ($K = 4$) nor a longer ($K = 16$) list falls far behind, so O^{3}PD is not sensitive to the exact choice of K within this range.

5 Analysis

Section 4.2 shows that O^{3}PD produces stronger students than OPD and, on the 8B→1.7B pair, keeps improving where OPD degrades. We now examine where this advantage comes from. We first ask what the student acquires from the teacher under ordinal supervision (Section 5.1), and then how training reshapes the student relative to its initialization (Section 5.2).

5.1 Learning the Teacher's Order, Not Its Probabilities

We measure token-level agreement between the student and the teacher on student-generated responses for the 8B→1.7B pair (Figure 5). At each position t, let $\mathcal{T}_t^T$ and $\mathcal{T}_t^S$ denote the top-8 tokens of the teacher and the student. *JSD@Top-8* is the Jensen–Shannon divergence between the two distributions restricted to $\mathcal{T}_t^T$ and renormalized, and measures how closely the student reproduces the teacher's probability values. *Kendall's τ* compares the orderings that the two models induce on $\mathcal{T}_t^T$, and *top-8 overlap* $|\mathcal{T}_t^T \cap \mathcal{T}_t^S|/8$ measures whether the student considers the same candidates plausible. *Top-1 agreement* is the fraction of positions at which the two models share the most likely token, and *student entropy* is the entropy of the student's next-token distribution over the full vocabulary. All metrics are averaged over positions.

The first metric concerns probability values. OPD behaves as its objective prescribes and keeps the JSD to the teacher low throughout training. The JSD of O^{3}PD is consistently higher and grows as training proceeds, which is expected since O^{3}PD never asks the student to match the teacher's probability values. On ordering and candidate set, the relation reverses. O^{3}PD achieves

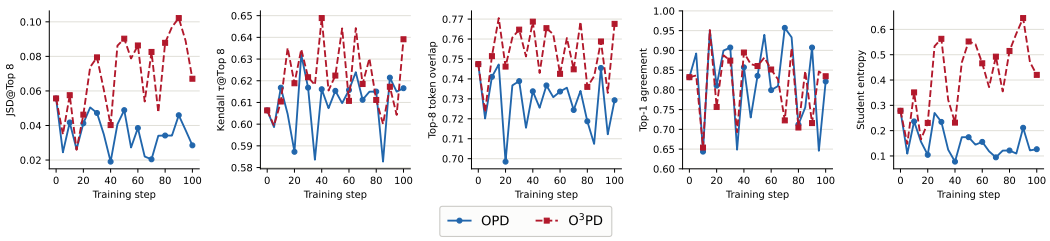

Figure 5: Token-level agreement between student and teacher during training for the 8B→1.7B pair, computed on student-generated responses. From left to right: JSD@Top-8, Kendall's τ, top-8 overlap, top-1 agreement, and student entropy. O^{3}PD departs further from the teacher's probabilities than OPD, yet agrees more closely with its ordering and candidate set, and maintains higher entropy throughout training.

a higher Kendall's τ than OPD for most of training and ends clearly above it, and its top-8 overlap stays above that of OPD at nearly every evaluation, while top-1 agreement is comparable. O^{3}PD thus transfers the structure that governs generation, namely which candidates the teacher considers plausible and how it orders them, and leaves the probability values free. The student also keeps a broader distribution: its entropy stays above that of OPD throughout training, whereas probability matching pulls the student toward the teacher's confidence profile. Together with Figure 3, these results show that closer probability matching does not translate into better reasoning: OPD attains the lower JSD, yet O^{3}PD attains the higher accuracy.

5.2 Reshaping the Student

Figure 4 places each checkpoint according to its distance from the base student and its distance from the teacher. For every student size, O^{3}PD checkpoints sit further from the teacher's probabilities than OPD checkpoints, consistent with Section 5.1. How far the student moves from its initialization, however, depends on student size. For the 1.7B student, OPD checkpoints stay close to the base student, whereas O^{3}PD checkpoints move several times further away. For the 4B student, the two methods move comparable distances, and for the 8B student, OPD moves further than O^{3}PD. Ordinal supervision thus does not uniformly produce larger updates; it produces them for the smallest student, where matching the teacher's probabilities is hardest.

The mechanism is intuitive. A small student cannot reproduce the probability values of a much larger teacher, so a magnitude-matching objective spends much of its capacity on a target that the student cannot reach. Ordinal supervision removes this constraint. The student is free to restructure its own distribution, provided that it preserves the teacher's ordering of the candidates, and the smallest student makes the most use of this freedom.

6 Conclusion

We presented Order-Only On-Policy Distillation (O^{3}PD), which trains the student to reproduce the teacher's top-K token ranking without matching the teacher's probability values. Because the supervision depends only on the teacher's order, O^{3}PD is invariant to the teacher temperature and requires no temperature tuning. Across six Qwen3 teacher–student configurations, O^{3}PD outperforms OPD on the three-benchmark average in every configuration, also exceeds SFT and GRPO, and, on the 8B→1.7B pair, keeps improving where OPD degrades. Our analysis shows that O^{3}PD reproduces the teacher's ranking and candidate set more faithfully than OPD while departing further from its probability values, and that it allows the smallest student to move substantially further from its initialization. The order-only principle is not tied to a single loss: other ranking objectives such as ListMLE also perform well. These results answer the question posed at the outset: for on-policy distillation, the teacher's ordering is sufficient, and matching its probability values is not only unnecessary but can hold the student back.

ETHICS STATEMENT

This work studies training objectives for knowledge distillation and does not involve human subjects, user data, or personally identifiable information. All experiments use publicly released models from the Qwen3 family and publicly available data: the mathematical reasoning subset of OpenThoughts for training, and the HMMT25, AIME24, and AIME25 competition problems for evaluation. We use these resources in accordance with their respective licenses and do not release any new datasets. Our experiments are restricted to mathematical reasoning, so we do not evaluate how ordinal distillation affects safety alignment or other behaviors in open-ended generation. We are not aware of any conflicts of interest related to this work.

REPRODUCIBILITY STATEMENT

We have made several efforts to ensure the reproducibility of our results. The O^3PD objective is fully specified in Section 3 (Equations (3) and (4)), and one training iteration is given as pseudocode in Algorithm 1. Appendix A describes the experimental setup: Table 5 lists all training and evaluation hyperparameters, including the LoRA configuration, rollout length, candidate-set size, decoding parameters, and random seeds, and Figure 6 gives the exact prompt template used for both training rollouts and evaluation. The OPD baseline shares the same setup. All models (Qwen3), training data (OpenThoughts), and evaluation benchmarks (HMMT25, AIME24, AIME25) are publicly available, and our evaluation protocol (Avg@12 over three seeds, answer extraction from `\boxed{}`) is described in Section A. We provide anonymized core source code for training in the supplementary material.

AI USE STATEMENT

We used generative AI tools to edit the paper for readability, to check the formatting of the manuscript and references, to search for related literature, and to assist in writing the code. All AI-assisted edits were reviewed by the authors, all suggested references were verified by the authors, and all AI-assisted code was checked by the authors. We take responsibility for the final content of this work.

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 B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/5be69a584901a26c521c2b51e40a4c20-Paper-Conference.pdf.

Ejafa Bassam, Dawei Zhu, and Kaigui Bian. Pld: A choice-theoretic list-wise knowledge distillation. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 136090–136112. Curran Associates, Inc., 2025. doi: 10.52202/085713-4542. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/c6b2921f24f82dc05cda53bbe50bab90-Paper-Conference.pdf.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4793–4801, 2019. doi: 10.1109/ICCV.2019.00489.

Wen-Shu Fan, Su Lu, Xin-Chun Li, De-Chuan Zhan, and Le Gan. Revisit the essence of distilling knowledge through calibration. In *ICML*, pp. 12882–12894, 2024.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/8ac015d409635f196f9e3e9dcfb9a94e-Paper-Conference.pdf.

Yuchen Guan, Runxi Cheng, Kang Liu, and Chun Yuan. Enhancing logits distillation with Plug&Play Kendall's τ ranking loss. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 20537–20552. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/guan25a.html.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer, Leon Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, Wanjia Zhao, John Yang, Shreyas Pimpalgaonkar, Kartik sharma, Charlie Ji, Yichuan Deng, Sarah Pratt, Vivek Ramanujan, Jon Saad-Falcon, Stutee Acharya, Jeffrey Li, Achal Dave, Alon Albalak, Kushal Arora, Blake Wulfe, Chinmay Hegde, Greg Durrett, Sewoong Oh, Mohit Bansal, Saadia Gabriel, Aditya Grover, Kai-Wei Chang, Vaishaal Shankar, Aaron Gokaslan, Mike Merrill, Tatsunori Hashimoto, Yejin Choi, Jenia Jitsev, Reinhard Heckel, Maheswaran Sathiamoorthy, Alex Dimakis, and Ludwig Schmidt. Openthoughts: Data recipes for reasoning models. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 108059–108130, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/b010241b9f1cdfc7d4c392db899cef86-Paper-Conference.pdf.

Changyi He, Yifu Ding, Jinyang Guo, Ruihao Gong, Haotong Qin, and Xianglong Liu. DA-KD: Difficulty-aware knowledge distillation for efficient large language models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 22379–22391. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/he25c.html.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.507. URL https://aclanthology.org/2023.findings-acl.507/.

Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 33716–33727. Curran Associates, Inc., 2022. doi: 10.52202/068431-2443. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/da669dfd3c36c93905a17ddba01eef06-Paper-Conference.pdf.

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

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards streamlined distillation for large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24872–24895. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/ko24c.html.

Jahyun Koo, Yerin Hwang, Yongil Kim, Taegwan Kang, Hyunkyung Bae, and Kyomin Jung. SWITCH: Studying with teacher for knowledge distillation of large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 3733–3746, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.206. URL https://aclanthology.org/2025.findings-naacl.206/.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention, 2023. URL https://arxiv.org/abs/2309.06180.

Cong Li and Gong Cheng. Order-preserving knowledge distillation. *Pattern Recognition*, pp. 114405, 2026.

Minchong Li, Feng Zhou, and Xiaohui Song. BiLD: Bi-directional logits difference loss for large language model distillation. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 1168–1182, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.78/.

Yongqi Li, Zhen Zhang, Wenjie Wang, Liqiang Nie, Wenjie Li, and Tat-Seng Chua. Distillation enhanced generative retrieval. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 11119–11129, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.662. URL https://aclanthology.org/2024.findings-acl.662/.

R Duncan Luce. *Individual choice behavior: A theoretical analysis*. Courier Corporation, 2005.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 5191–5198, 2020.

Robin L Plackett. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2):193–202, 1975.

Sashank Reddi, Rama Kumar Pasumarthi, Aditya Menon, Ankit Singh Rawat, Felix Yu, Seungyeon Kim, Andreas Veit, and Sanjiv Kumar. Rankdistil: Knowledge distillation for ranking. In *International Conference on Artificial Intelligence and Statistics*, pp. 2368–2376. PMLR, 2021.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. Logit standardization in knowledge distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15731–15740. IEEE, 2024.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45, 2020.

Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. Listwise approach to learning to rank: theory and algorithm. In *Proceedings of the 25th international conference on Machine learning*, pp. 1192–1199, 2008.

Wenda Xu, Rujun Han, Zifeng Wang, Long Le, Dhruv Madeka, Lei Li, William Yang Wang, Rishabh Agarwal, Chen-Yu Lee, and Tomas Pfister. Speculative knowledge distillation: Bridging the teacher-student gap through interleaved sampling. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=EgJhwYR2tB.

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

Mengxiang Zhang and Lingyuan Liu. Staged knowledge distillation through least-to-most prompting: Optimizing teacher guidance via difficulty-aware training. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 8489–8501, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.451. URL https://aclanthology.org/2025.findings-emnlp.451/.

Zhengze Zhang, Shiqi Wang, Yiqun Shen, Simin Guo, Dahua Lin, Xiaoliang Wang, Cam Tu Nguyen, and Fei Tan. daDPO: Distribution-aware DPO for distilling conversational abilities. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 15421–15437, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.796. URL https://aclanthology.org/2025.findings-acl.796/.

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. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Jpxfof0EaS.

A Experimental Setup

A.1 Hyperparameters

Table 5 lists the hyperparameters used for training and evaluation. We fine-tune the student with LoRA and train for 100 steps with an effective batch size of 32. During training, the teacher runs in thinking mode while the student runs in non-thinking mode, and each student rollout is limited to 1,024 tokens. At every student-generated prefix, both OPD and O^{3}PD use the teacher's top-8 candidates. For evaluation, we checkpoint every 20 steps and report Avg@12, obtained by drawing four samples per problem under each of three seeds.

A.2 Prompt Template

All training rollouts and evaluations use the same prompt, shown in Figure 6. The prompt asks the model to reason step by step and to place its final answer in \boxed{}, from which we extract the answer for scoring.

Prompt template

```
Problem: {problem}
Reason step by step and put the final answer in \boxed{}.
```

Figure 6: Prompt template used for training rollouts and evaluation. {*problem*} is replaced with the problem statement.

Table 5: Hyperparameters for O^{3}PD and OPD training and evaluation. All methods share the same settings unless noted otherwise.

Hyperparameter	**Value**
Training	
Student fine-tuning	LoRA ($r = 64$, $\alpha = 128$)
Learning Rate	5e-6
Teacher thinking mode	Enabled
Student thinking mode	Disabled
Rollout length	1,024 tokens
Effective batch size	32
Training steps	100
Training seed	42
Top-K candidates (OPD and O^{3}PD)	8
Ranking objective (O^{3}PD)	Adjacent logistic
Teacher temperature (OPD and O^{3}PD)	1
Evaluation	
Evaluation interval	Every 20 steps
Evaluation seeds	41, 42, 43
Samples per problem	4 per seed (12 total, Avg@12)
Sampling temperature	1.0
Top-p	1.0
Top-k	-1 (disabled)
Min-p	0
Presence penalty	0
Max new tokens	38,912
Max model length	40,960

Algorithm 1 Order-Only On-Policy Distillation (O^{3}PD)

Require: Prompt set $\mathcal{D}$; frozen teacher p^T; student policy π_θ; top-K size K

1: **while** not converged **do**

2: Sample a minibatch $\mathcal{B} \subset \mathcal{D}$

3: **for all** $x \in \mathcal{B}$ **do**

4: Sample an on-policy response $y^S \sim \pi_\theta(\cdot \mid x)$

5: At each $s_t = (x, y^S_{<t})$, obtain

$$(v_{t,(1)}, \ldots, v_{t,(K)}) \leftarrow \mathrm{TopKOrder}\big(p_t^T(\cdot)\big)$$

6: Compute the adjacent ranking loss

$$\ell(x) \leftarrow -\frac{1}{|y^S|(K-1)} \sum_{t=1}^{|y^S|} \sum_{i=1}^{K-1} \log \frac{p_t^S(v_{t,(i)})}{p_t^S(v_{t,(i)}) + p_t^S(v_{t,(i+1)})}$$

7: $\mathcal{L}_{\mathrm{adj}} \leftarrow \dfrac{1}{|\mathcal{B}|} \displaystyle\sum_{x \in \mathcal{B}} \ell(x)$; update θ

A.3 Training Procedure

Algorithm 1 summarizes one training iteration of O^{3}PD. For each prompt, the student samples a response, and the teacher is queried at every prefix of that response to obtain its top-K tokens in descending order of probability. The student is then updated with the adjacent ranking loss, which averages the pairwise logistic terms over all positions and all $K - 1$ adjacent pairs. The teacher's probabilities are used only to select and order the candidates and do not appear in the loss.

A.4 Infrastructure

All models are trained with Transformers Wolf et al. (2020) and evaluated using vLLM Kwon et al. (2023) on NVIDIA H100 GPUs.

B Teacher Temperature Sensitivity of OPD

Figure 7 reports the teacher-temperature sweep. We train standard OPD for every teacher–student pair at each $T \in \{0.25, 0.5, 1, 2, 4, 8\}$ and evaluate HMMT25 Avg@12. The sweep covers the six configurations in Table 3 and additionally includes a Qwen3-0.6B student.

OPD is highly sensitive to T, and no single temperature is best across pairs: the optimum ranges from $T = 0.25$ (8B→1.7B) to $T = 2$ (14B→8B, 4B→1.7B). Sensitivity increases as the student becomes smaller. For the 0.6B student, performance collapses at high temperatures, e.g., 14B→0.6B falls from 11.94 at $T = 1$ to 1.11 at $T = 8$. O^{3}PD avoids this selection problem entirely, since its ranking targets are identical for every $T > 0$.

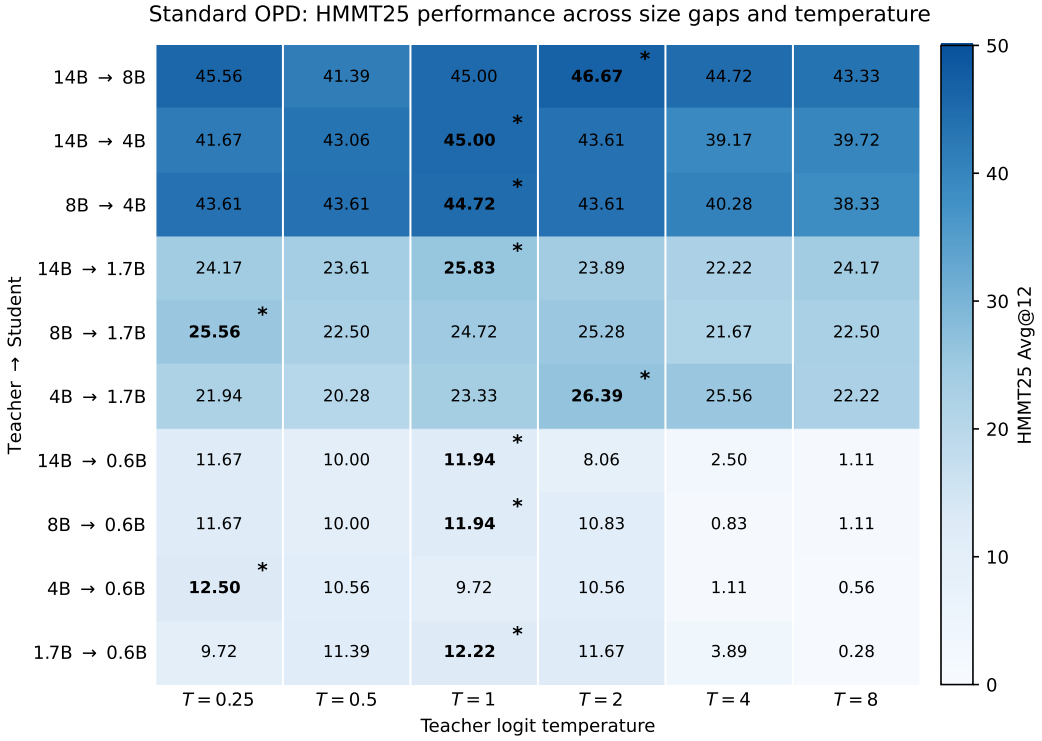

Figure 7: HMMT25 Avg@12 of standard OPD across teacher–student pairs (rows) and teacher logit temperatures T (columns). Each run is trained for only 3 steps with a batch size of 32. Bold values marked with $*$ indicate the best temperature for each pair.