BETTER SUPERVISION IS NEARBY: NEIGHBORHOOD ON-POLICY SELF-DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy self-distillation (OPSD) trains mathematical reasoning models using a privileged teacher that sees a reference solution and supervises student-sampled prefixes. Standard OPSD uses one fixed parameter setting at every state, but nearby settings may offer additional supervision. We find that local parameter perturbations reveal complementary reference-aligned corrections under the same reference context. Different experts supply these corrections at different reference positions. Their pool covers more such positions than the unperturbed privileged teacher. We introduce NEIGHBORHOOD OPSD (N-OPSD) to turn these corrections into supervision at student-visited states. Offline, greedy selection builds a compact pool of frozen experts by rewarding filtered reference-token gains beyond the pool's current best at each position. The highest-peak expert need not provide the best training target. Online routing therefore separates the anchor direction from its level of support. MaxPeak selects the anchor token, and quantile selection chooses among experts whose top token matches it. The student learns from the chosen expert's full next-token distribution through the clipped forward-KL objective inherited from OPSD. We evaluate on AIME 2024, AIME 2025, and HMMT February 2025. Across three independent runs per method, NEIGHBORHOOD OPSD improves the three-benchmark Average@12 over OPSD by 2.75, 1.67, and 1.94 points on Qwen3-1.7B, 4B, and 8B, respectively. Student-prefix continuations support using the pool beyond the reference trajectories used for selection. Matched ablations support filtered reference-token gains as a selection criterion. Accounting for overlap within the pool and routing by state further improve student accuracy. Inference uses only the distilled student.

1 INTRODUCTION

On-policy self-distillation (OPSD) trains mathematical reasoning models using two copies of the same model (Zhao et al., 2026). The student generates a response from the problem alone. The privileged teacher also sees a reference solution, available only during training (Lopez-Paz et al., 2016), and provides next-token distributions along the student's response. Standard OPSD uses the same unperturbed teacher M_0 at every student-visited state. Its predictions depend on the student prefix, but all supervision comes from one fixed parameter setting. A fixed teacher need not exhaust the useful supervision available from the same base model. We study its local parameter neighborhood as a source of complementary corrections for on-policy distillation.

TrustMOPD weights specialist supervision at each student prefix (Sun et al., 2026). RandOpt selects local perturbation experts by task performance (Gan & Isola, 2026). For distillation, an expert's task accuracy need not reflect the value of its supervision (Cho & Hariharan, 2019). Task accuracy also does not reveal which corrections the selected experts already supply. This motivates selecting local experts by the token-level corrections they add to a pool.

Experts sharing the same base checkpoint and reference context provide complementary corrections in our reference-side analysis. We measure each correction as an increase in reference-token probability over the frozen problem-only model on the same reference prefix. Different experts provide these increases at different positions. Their pool covers more reference positions than the unperturbed privileged teacher M_0 under the same selection filters. Thus, useful supervision can depend on which nearby parameter setting interprets the same privileged information.

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

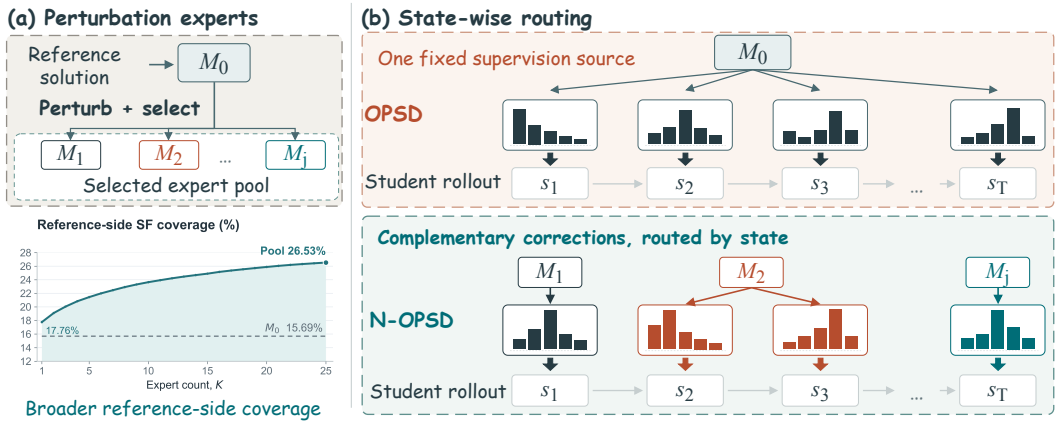


Figure 1: Overview of NEIGHBORHOOD OPSD. (a) Selected local perturbation experts expand reference-side selection-feasible (SF) coverage beyond M_0 (Qwen3-8B, $\sigma = .002$; 500-question analysis set). (b) OPSD uses M_0 at every student-visited state, while NEIGHBORHOOD OPSD routes one expert's full next-token distribution at each state. Colors distinguish selected experts; distributions are schematic.

We introduce NEIGHBORHOOD OPSD (N-OPSD) to turn complementary corrections in the teacher's neighborhood into supervision at student-visited states (Figure 1). Offline selection builds a compact pool by rewarding experts for reference-aligned corrections beyond what the current pool provides. During training, student prefixes can differ from the reference solution (Agarwal et al., 2024), so the reference token cannot directly guide routing. The selected experts therefore evaluate the student's current prefix.

Online routing separates the anchor direction from its level of support. MaxPeak takes the token with the highest probability assigned by any expert as the anchor, allowing an individual expert to supply the direction without pool agreement. A large positive gap between teacher and student probabilities can cause the anchor's loss term to be clipped (Zhao et al., 2026). Quantile selection chooses the level of support among experts whose top token matches the anchor. The student learns from the chosen expert's full next-token distribution. All experts remain fixed, and inference uses only the distilled student.

We evaluate on AIME 2024, AIME 2025, and HMMT February 2025. Across three independent runs per method, NEIGHBORHOOD OPSD improves the three-benchmark Average@12 over OPSD by 2.75, 1.67, and 1.94 points on Qwen3-1.7B, 4B, and 8B. Student-prefix continuations also show that MaxPeak can improve continuation accuracy over M_0. This supports using the selected pool beyond the reference trajectories used for expert selection. Selection ablations show gains from scoring reference-aligned corrections, with further gains from accounting for overlap within the pool. Routing ablations then show gains from choosing the supervision source based on expert predictions at each state.

The contributions are:

- We show that a privileged teacher's reference-aligned supervision extends beyond its unperturbed parameter setting. Selected local experts provide complementary corrections at reference positions beyond those covered by M_0.
- We introduce NEIGHBORHOOD OPSD, a framework that turns a privileged teacher's neighborhood into state-wise supervision for on-policy distillation. It selects complementary local experts and distills their supervision into a single student used alone at inference.
- Across three independent runs per method, NEIGHBORHOOD OPSD improves the three-benchmark Average@12 over OPSD by 1.67 to 2.75 points on Qwen3-1.7B, 4B, and 8B. Matched ablations connect these gains to complementary expert selection and state-wise routing.

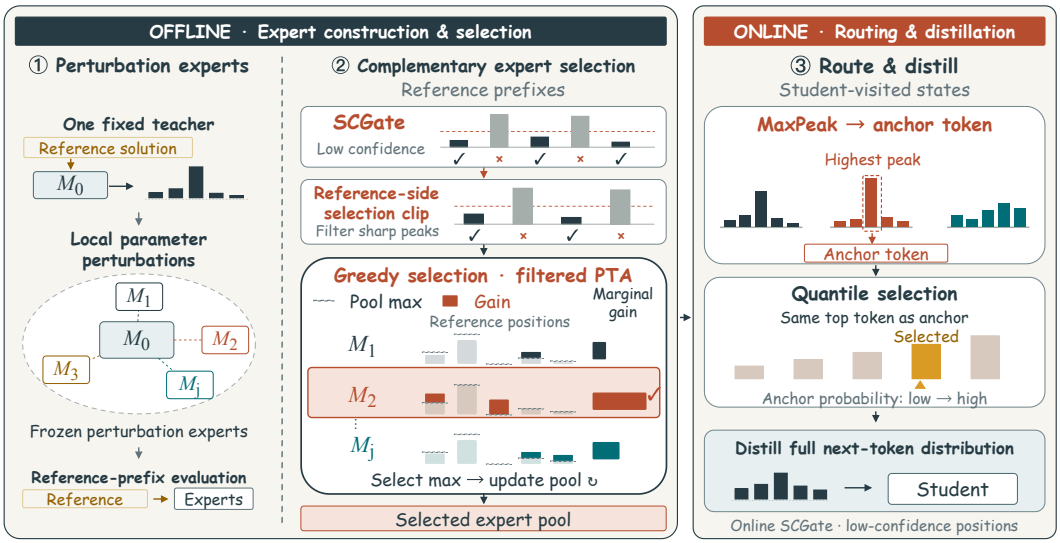


Figure 2: Overview of NEIGHBORHOOD OPSD. Offline, greedy selection builds a complementary pool of local perturbation experts by marginal filtered PTA on reference prefixes. Online, MaxPeak selects an anchor token, and quantile selection chooses among experts whose top token matches it. The selected expert supplies the full next-token distribution for student distillation.

2 METHOD

NEIGHBORHOOD OPSD separates the choice of an expert pool from the choice of a supervision source at each state (Figure 2). Offline selection identifies experts that add complementary corrections on reference prefixes. Online routing then selects from this pool at the student's current prefix, where the useful supervision source can change as the rollout develops.

2.1 SETUP AND PERTURBATION EXPERTS

Let x be a problem, z a privileged reference solution available only during training, and π_θ the student. The student samples a response $\mathbf{y} = (y_1, \ldots, y_T)$ from the problem-only prompt:

$$\mathbf{y} \sim \pi_\theta(\cdot \mid x), \qquad s_t = (x, y_{<t}). \tag{1}$$

At each student-visited state s_t, the unperturbed privileged teacher M_0 has parameters $\bar{\theta}$ and observes both z and the same student prefix:

$$p_t(v) = \pi_\theta(v \mid s_t), \qquad q_t^0(v) = \pi_{\bar{\theta}}(v \mid x, z, y_{<t}). \tag{2}$$

The teacher returns a next-token distribution along the student rollout. OPSD uses q_t^0 at every state and applies a pointwise forward-KL clip with threshold κ (Zhao et al., 2026). NEIGHBORHOOD OPSD retains this clip and selects the supervision source from an expert pool.

For seed r_j and radius σ, perturbation expert $j \geq 1$ has parameters

$$\bar{\theta}_j = \bar{\theta} + \epsilon_j, \qquad \epsilon_j \sim \mathcal{N}(0, \sigma^2 I). \tag{3}$$

The noise is deterministic for each (r_j, σ). The candidate pool includes these perturbation experts and M_0, with $\bar{\theta}_0 = \bar{\theta}$. All experts receive the same privileged information and student prefix, so their diversity comes from the local parameter neighborhood. The teacher snapshot and all experts remain fixed throughout training. We select a compact expert pool offline and route one expert at each student-visited state. The radius is calibrated with student-prefix continuations before downstream evaluation.

2.2 COMPLEMENTARY EXPERT SELECTION

We select experts for the corrections they add to the pool. At each reference position, we credit a candidate only for its filtered correction beyond the pool's current best.

Candidate selection uses teacher forcing on reference solutions. Let y_t^* be the reference token. At position t, the fixed problem-only model and every candidate expert evaluate the same reference prefix:

$$p_t^S(v) = \pi_{\bar{\theta}}(v \mid x, y_{<t}^*), \qquad q_t^j(v) = \pi_{\bar{\theta}_j}(v \mid x, z, y_{<t}^*). \tag{4}$$

The shared prefixes make candidate corrections comparable at the same state and reference token. The fixed problem-only model provides the same baseline for every candidate, including M_0.

To focus selection on lower-confidence reference positions, we apply a student-confidence gate (SC-Gate):

$$g_t^{\mathrm{ref}} = \mathbb{I}[p_t^S(y_t^*) \le \tau_{\mathrm{sel}}]. \tag{5}$$

For each retained position, we compute the reference-token forward-KL contribution

$$c_{j,t} = q_t^j(y_t^*) \log \frac{q_t^j(y_t^*)}{p_t^S(y_t^*)}. \tag{6}$$

We set $\kappa_{\mathrm{sel}} = \kappa$ and give no selection credit to pairs with $c_{j,t} > \kappa_{\mathrm{sel}}$. This reference-side selection clip filters sharp reference-token peaks that the clipped online objective may suppress. It examines the reference token on a fixed reference prefix, while the online clip acts on every vocabulary entry at student-visited states.

Positive teacher advantage (PTA) measures how much an expert raises the reference-token probability. Applying both filters gives filtered PTA:

$$\begin{aligned} A_{j,t} &= \left[q_t^j(y_t^*) - p_t^S(y_t^*) \right]_+, \\ \widetilde{A}_{j,t} &= g_t^{\mathrm{ref}} \mathbb{I}[c_{j,t} \le \kappa_{\mathrm{sel}}] A_{j,t}. \end{aligned} \tag{7}$$

A positive $\widetilde{A}_{j,t}$ denotes a reference-aligned, selection-feasible correction. For a subset S of candidate experts, we maximize

$$\mathrm{Score}(S) = \sum_{(x,z)} \sum_t \max_{j \in S} \widetilde{A}_{j,t}. \tag{8}$$

The score rewards both newly covered positions and stronger corrections at positions already covered. It therefore accounts for correction magnitudes, whereas selection-feasible coverage counts positions with positive filtered PTA. If two experts supply similar corrections at the same positions, selecting one reduces the other's marginal gain. An expert that improves different positions can then become more useful to the pool. Starting from $S = \varnothing$, with the empty-set maximum defined as zero, we add the remaining expert that increases $\mathrm{Score}(S)$ the most. Selection continues until K experts are chosen. In our experiments, the PTA selector does not choose M_0 within the first 50 greedy steps.

2.3 MaxPeak-Anchored Quantile Routing

Offline selection can reward an expert for a correction that no other expert supplies. Choosing the anchor by pool consensus could suppress such an individual signal. We therefore use maximum expert support to choose the online anchor direction. At each student-visited state, expert $j \in S$ produces $q_t^j = \pi_{\bar{\theta}_j}(\cdot \mid x, z, y_{<t})$. Since a student prefix need not have an aligned reference token, we use MaxPeak to select the highest-peak expert and take its top token as the anchor:

$$j_t^* = \arg\max_j \max_{v \in \mathcal{V}} q_t^j(v), \qquad a_t^* = \arg\max_v q_t^{j_t^*}(v). \tag{9}$$

This selects the token with the largest probability assigned by any expert.

We then collect experts whose top token matches the anchor:

$$\mathcal{E}_t = \left\{ j \in S : \arg\max_v q_t^j(v) = a_t^* \right\}. \tag{10}$$

The MaxPeak expert j_t^* belongs to this set. For each $j \in \mathcal{E}_t$, the anchor probability gap is $A_{j,t}^{\mathrm{on}} = q_t^j(a_t^*) - p_t(a_t^*)$. At a fixed state, ranking these gaps is equivalent to ranking the experts' anchor

probabilities. The eligible experts share a top token, but their full next-token distributions can still differ.

A large positive gap can cause the anchor's forward-KL contribution to be clipped, removing that term's gradient. A lower anchor probability may offer little or no increase over the student's probability. To choose the level of anchor support, we sort the gaps in ascending order and select the expert at the lower discrete quantile:

$$k_t = \lfloor q\,(|\mathcal{E}_t| - 1) \rfloor\,, \qquad \widehat{j}_t = \mathrm{order}_{\mathcal{E}_t}(k_t). \tag{11}$$

Here, $\mathrm{order}_{\mathcal{E}_t}$ indexes the sorted experts from zero. The endpoints $q = 0$ and $q = 1$ select the smallest and largest anchor probability gaps, respectively. Thus, $q = 1$ selects a highest-peak expert. The selected expert supplies the full-vocabulary target $\widehat{q}_t = q_t^{\widehat{j}_t}$; the anchor token is used only for routing. The anchor and matching set are recomputed at each state, so routing can select different experts as the student visits new prefixes.

2.4 Training Objective

The online SCGate retains positions where the student assigns probability at most τ to its sampled token: $g_t = \mathbb{I}[p_t(y_t) \le \tau]$. For each retained position, we apply the inherited OPSD pointwise clip to every vocabulary-level forward-KL contribution:

$$\ell_{t,v} = \widehat{q}_t(v) \log \frac{\widehat{q}_t(v)}{p_t(v)}, \qquad \overline{\ell}_{t,v} = \min(\ell_{t,v}, \kappa). \tag{12}$$

The training loss is

$$\mathcal{L}_{\text{N-OPSD}} = \frac{1}{\sum_t g_t} \sum_t g_t \sum_{v \in \mathcal{V}} \overline{\ell}_{t,v}. \tag{13}$$

The denominator counts SCGate-retained positions. The pointwise clip is applied before summing over the full vocabulary. Contributions below κ remain unchanged. Only the student parameters θ receive gradients. Training repeats this process on rollouts sampled from the current student. The pool remains fixed, but its predictions and the routed targets are recomputed on each rollout. Appendix A details the selection clip, greedy marginal gain, and equivalent token-support form of MaxPeak.

3 Experiments

3.1 Experimental Setup

Models and training. We train Qwen3-1.7B, 4B, and 8B (Yang et al., 2025) on 10,000 OpenThoughts problems (Guha et al., 2025) for 100 optimizer steps. At each scale, the student and frozen experts start from the same base checkpoint. Unless varied, Neighborhood OPSD uses $K = 25$ experts and routing quantile $q = .75$.

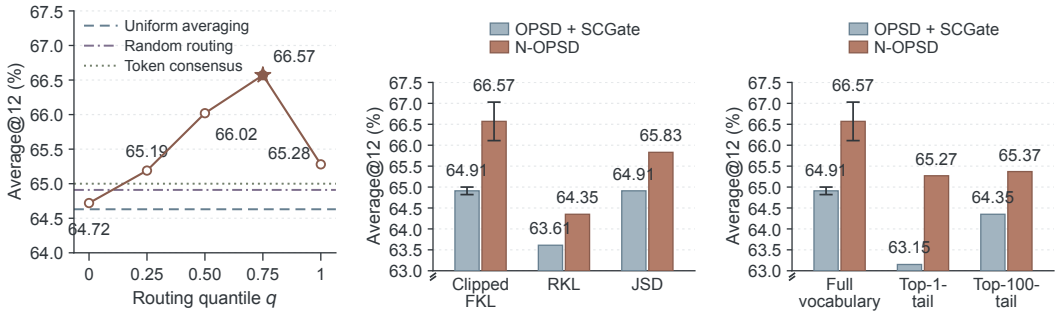
Baselines. We use the Base, supervised fine-tuning (SFT), and GRPO scores from Zhao et al. (2026) and rerun OPSD, EOPD, and PW-OPSD (Zhao et al., 2026; Jin et al., 2026; Liu et al., 2026a) under the same initialization, data, training budget, and evaluation as Neighborhood OPSD. OPSD+SCGate isolates the online gate's effect with the fixed teacher M_0.

Benchmarks and metrics. We evaluate AIME 2024, AIME 2025, and HMMT February 2025 (Mathematical Association of America, 2024; 2025; Harvard-MIT Mathematics Tournament, 2025) using the OPSD sampling settings (Zhao et al., 2026), reporting Average@12 from 12 sampled responses per problem. Avg. is the unweighted three-benchmark mean. Following PW-OPSD (Liu et al., 2026a), we report the step-100 checkpoint for Neighborhood OPSD and all rerun baselines. Neighborhood OPSD, OPSD, EOPD, and PW-OPSD use three independent runs per scale; the 8B OPSD+SCGate control also uses three runs. We report means and sample standard deviations for these methods. Other results are point estimates unless noted. Appendix B gives all configurations and implementation details.

Table 1: Main results across three Qwen3 scales with Average@12 accuracy (%; higher is better).

Model	Metric	Base†	SFT†	GRPO†	OPSD	EOPD	PW-OPSD	N-OPSD
Qwen3 1.7B	AIME24	51.50	48.40	51.10	*55.46* ± 0.42	52.41 ± 0.64	57.32 ± 0.70	**57.96** ± 0.42
	AIME25	36.70	36.30	38.30	*40.93* ± 0.70	39.35 ± 0.42	41.76 ± 0.70	**45.09** ± 0.97
	HMMT25	23.10	22.70	23.70	*29.26* ± 0.42	25.93 ± 0.70	30.46 ± 0.58	**30.83** ± 1.11
	Avg.	37.10	35.80	37.70	*41.88* ± 0.19	39.23 ± 0.23	43.18 ± 0.14	**44.63** ± 0.56
Qwen3 4B	AIME24	74.90	70.20	75.60	*76.20* ± 0.16	73.43 ± 0.58	76.48 ± 0.42	**76.95** ± 0.74
	AIME25	66.40	62.30	*68.10*	67.87 ± 0.70	66.11 ± 1.00	68.70 ± 0.85	**70.83** ± 0.56
	HMMT25	42.20	43.40	*44.40*	45.55 ± 0.74	42.32 ± 0.85	42.68 ± 0.85	**46.85** ± 0.70
	Avg.	61.17	58.63	*62.70*	63.21 ± 0.11	60.62 ± 0.46	62.62 ± 0.14	**64.88** ± 0.39
Qwen3 8B	AIME24	75.80	72.30	76.40	77.41 ± 0.32	*77.50* ± 0.56	77.59 ± 0.64	**78.70** ± 0.69
	AIME25	65.60	64.20	68.90	71.02 ± 0.32	70.92 ± 0.97	*70.93* ± 0.70	**72.78** ± 0.28
	HMMT25	43.90	42.90	*46.70*	45.46 ± 0.42	45.46 ± 0.64	46.85 ± 0.80	**48.24** ± 0.89
	Avg.	61.77	59.80	64.00	*64.63* ± 0.16	*64.63* ± 0.09	65.12 ± 0.33	**66.57** ± 0.46

†: results from Zhao et al. (2026); other entries are our means ± sample standard deviations over three independent training runs. **Bold**/underlined/*italic*: first/second/third in each row, with gold shading from darkest to lightest; ties share a rank. Avg.: unweighted mean across the three benchmarks.



(a) Quantile routing (b) Online divergence (c) Target distribution

Figure 3: Ablations on Qwen3-8B. Values are three-benchmark Average@12 (%). Error bars show means ± sample standard deviations over three training runs; results without error bars are single runs.

3.2 Main Results Across Model Scales

Neighborhood OPSD ranks first across all nine model-benchmark settings (Table 1). Its three-benchmark average exceeds OPSD by 2.75, 1.67, and 1.94 points on 1.7B, 4B, and 8B, respectively. Downstream ablations use Qwen3-8B and report single-run three-benchmark Average@12 unless noted. Appendix C gives per-benchmark ablation results, matched single-run comparisons, and details of target compression.

3.3 Complementary Expert Selection

Reference-side coverage. On 500 training questions held out from selection, the pool covers 26.53% of reference-side SCGate-retained positions at $\sigma = .002$, compared with 15.69% for M_0. Coverage requires positive filtered PTA from at least one expert. The pool has higher coverage at every tested radius (Table 2). Coverage also increases as experts are added (Appendix D.3), supporting complementarity within the selected pool. With reference context shared, this additional coverage comes from variation in teacher parameters. Newly covered reference tokens at $\sigma = .002$ include common words, mathematical symbols, numbers, and verbs such as `find` and `need`. Appendix E reports the full reference-side audit and student-prefix continuation probe.

Table 2: Reference-side corrections, student-prefix continuations, and downstream accuracy across perturbation radii on Qwen3-8B (%).

Setting	Reference prefixes		Student-prefix continuations		Distilled student
	SF coverage	Top-1 acc.	Accuracy	Completion	Avg.
M_0	15.69	91.33	83.33	90.00	n/a
$\sigma = .001$	21.86	94.20	81.33	88.67	65.83
$\sigma = .002$	26.53	95.58	**86.67**	**94.67**	**66.57**
$\sigma = .004$	29.02	96.53	62.33	79.00	64.54
$\sigma = .006$	**31.39**	**97.31**	27.33	39.33	63.24

Reference metrics use SCGate-retained positions on 500 held-out questions. For each radius, selection-feasible (SF) coverage uses the pool and top-1 accuracy uses the MaxPeak anchor. Continuations use the same 300 student prefixes, privileged references, and 4096-token budget; all prefixes remain in the denominator. Continuation results for each radius use Raw MaxPeak without quantile selection. Avg. uses NEIGHBORHOOD OPSD with $K = 25$, $q = .75$, and matched single training runs.

Table 3: Expert selection and filtering on Qwen3-8B (three-benchmark Average@12, %).

Expert selection				Filtering		Fixed-teacher controls		Full method
Random Top-K	Quality Top-K	Greedy sample coverage	Independent PTA Top-K	w/o SCGate	w/o selection clip	OPSD	OPSD +SCGate	**Ours**
$64.29^{\pm.53}$ -2.28	65.00 -1.57	65.37 -1.20	65.65 -0.92	65.28 -1.29	64.26 -2.31	$64.63^{\pm.16}$ -1.94	$64.91^{\pm.09}$ -1.66	**66.57**$^{\pm.46}$

Small lower-right values show differences from the full method in percentage points. Its three-run mean and matched single-run score both equal 66.57. Superscripts give sample standard deviations over three training runs (Random Top-K: three random pools); other entries are single runs. SCGate removal applies to both stages. The online OPSD clip is retained in both component ablations.

Selection criteria. Table 3 tests how selection affects student accuracy. All selectors share 1,000 selection questions, 501 candidates, $\sigma = .002$, $K = 25$, and the same online router and gate. Quality Top-K ranks experts by answer accuracy, following RandOpt (Gan & Isola, 2026). Greedy sample coverage adds the expert that correctly answers the most additional questions.

Independent PTA Top-K ranks experts by their total filtered PTA. Like Quality Top-K, it ranks experts independently, but uses filtered reference-token corrections as the selection criterion. It reaches 65.65, exceeding Quality Top-K by $.65$ points. Our PTA selector uses greedy marginal gain and reaches 66.57, a further $.92$-point gain. Both PTA selectors use the same per-token filtered PTA values. They differ in whether selection accounts for corrections already supplied by the current pool. Both comparisons show gains on all three benchmarks (Table 4). Greedy sample coverage also rewards nonredundant contributions, but measures them as newly solved questions. Our selector exceeds its 65.37 average by 1.20 points. With routing held fixed, these results support filtered PTA as a selection criterion and the added value of selecting complementary corrections.

3.4 ROUTING ON STUDENT PREFIXES

Offline selection uses reference prefixes, while training visits prefixes sampled by the student. We first test the pool on student-prefix continuations, then examine how to choose its training targets.

Student-prefix continuations. MaxPeak reaches 95.58% reference-side anchor accuracy at $\sigma = .002$, versus 91.33% for M_0 on the same SCGate-retained positions. We test transfer using prefixes from the frozen base student on 300 distinct questions. These come from 150 completed wrong responses and 150 unfinished responses. Raw MaxPeak selects the highest-peak expert at each decoding step without quantile selection. Both Raw MaxPeak and M_0 use the same prefixes and

privileged references, with a budget of 4096 new tokens. Raw MaxPeak raises continuation accuracy from 83.33% to 86.67% and completion from 90.00% to 94.67%.

Table 2 connects these continuation results to downstream accuracy across radii. Both continuation accuracy and downstream Average@12 peak at $\sigma = .002$, although reference-side coverage and anchor accuracy continue to rise at larger radii. Higher reference-side coverage therefore need not yield better student supervision. This supports calibrating the radius on student prefixes.

Training-time routing. At $\sigma = .002$, Figure 3(a) compares training-time routing on the same PTA-selected pool. Uniform averaging takes the mean of all expert distributions. Token consensus equally averages the distributions of experts voting for the most common top token. Random routing samples one expert uniformly at each state and uses its full distribution. All rules share the same SCGate loss normalization. At $q = .75$, MaxPeak-anchored quantile routing gains 1.94 points over Uniform averaging and 1.57 over Token consensus. Our router raises Average@12 from 64.91 with Random routing to 66.57. Both routers use the same pool and supply one expert's full distribution at each state. This gain supports choosing that expert using the current expert predictions.

The quantile comparison fixes the pool and MaxPeak anchor rule while varying which anchor-consistent expert supplies the full distribution. Accuracy rises from $q = 0$ to .75 but drops 1.29 points at $q = 1$. Thus, the highest-peak expert need not provide the best training target, supporting separate choices of anchor direction and support level.

3.5 Distillation and Filtering

Online divergence. With the pool and router fixed, NEIGHBORHOOD OPSD outperforms OPSD+SCGate under both Jensen-Shannon divergence (JSD) and reverse KL (RKL) (Figure 3(b)), with gains on all three benchmarks. The source of supervision therefore matters beyond the default forward-KL objective. The clipped forward-KL reference has the highest average for NEIGHBORHOOD OPSD, with a three-run mean of 66.57.

Target distribution. Top-N-tail retains the teacher's N most probable tokens and combines the rest into one tail category. The student uses the same categories. For Top-1-tail, supervision is a binary soft target over the top token and the remaining mass. We compress the distributions after routing and then apply forward-KL clipping. NEIGHBORHOOD OPSD retains an average advantage for both Top-1-tail and Top-100-tail (Figure 3(c)), with a 2.12-point gain over OPSD+SCGate even for the binary target. Its full-vocabulary target scores 1.30 and 1.20 points higher than these two compressed variants, respectively.

SCGate and the selection clip. We also test which positions and reference-token gains to retain. In the reference-prefix audit, the problem-only model's top-1 confidence is at least .99 at 65.63% of positions. Its top-1 accuracy on these positions is 97.67%, motivating supervision of lower-confidence positions. With the fixed M_0 teacher, SCGate raises the three-run average from 64.63 to 64.91. NEIGHBORHOOD OPSD reaches 66.57, a further 1.66-point gain (Table 3). The online gate thus accounts for only part of the improvement over OPSD.

Removing both offline and online SCGate and reselecting the pool gives 65.28. Reselecting without the reference-side selection clip gives 64.26, with both SCGate stages and the online OPSD clip retained. This drop supports filtering expert credit during offline selection even when the training objective already clips the loss. Matched single-run comparisons in Appendix C give the same ordering. The default $\tau = .99$ gives the highest Average@12 among the four tested thresholds; Appendix D.2 reports the complete confidence audit and threshold sensitivity results.

3.6 Pool Size and Training Cost

At $\sigma = .002$ and $q = .75$, increasing the pool from $K = 1$ to 25 raises Average@12 from 64.54 to 66.57 (Appendix D.3). On the 500-question analysis set, $K = 25$ retains 94.81% of the selection-feasible coverage at $K = 50$.

Expanding to $K = 50$ adds reference-side coverage, but Average@12 falls to 64.35 at $q = .75$. Lowering q to .50 raises it to 65.83, still below the $K = 25$ result. This shift in the best tested

quantile supports choosing pool size together with the level of routed support. Among the tested settings, $K = 25$ gives the highest three-benchmark average.

Using $K = 25$ instead of $K = 50$ reduces estimated step time by 25.63%, to $1.423\times$ that of OPSD+SCGate. This estimate combines measured teacher time with shared rollout and student computation times (Appendix D.4). Inference uses only the distilled student.

4 Related Work

4.1 On-Policy and Privileged Self-Distillation

On-policy distillation supervises student-generated prefixes to reduce the state mismatch of off-policy imitation (Ross et al., 2011; Agarwal et al., 2024; Gu et al., 2024; Ko et al., 2024). Self-distillation can condition the teacher on demonstrations, feedback, or guiding context while training on student rollouts (Shenfeld et al., 2026; Hübotter et al., 2026; Ye et al., 2026). OPSD conditions the teacher on a reference solution unavailable to the student (Zhao et al., 2026). Tian et al. (2026) study the choice of privileged context, including context selection per trajectory. NEIGHBORHOOD OPSD keeps the reference context shared across experts and varies the teacher parameters.

4.2 Adapting Supervision from a Fixed Teacher

Fixed-teacher methods adapt the target, divergence, or loss weights. TAID gradually interpolates student and teacher logits (Shing et al., 2025), and DistiLLM-2 uses different divergences for teacher- and student-generated responses (Ko et al., 2025). EOPD adapts divergence to teacher entropy (Jin et al., 2026), while PW-OPSD and IW-OPD reweight token losses (Liu et al., 2026a; Xie et al., 2026). Other methods use lookahead feedback, trajectory weighting, hidden-state transitions, or purified targets (Liu et al., 2026b; Lin et al., 2026; Li et al., 2026; Shen et al., 2026). Cal-OPD calibrates token-level discrepancies using contrasting teacher contexts (He et al., 2026). NEIGHBORHOOD OPSD expands the supervision source through complementary local experts. Its routed targets can be used with divergence adaptation or loss weighting.

4.3 Multiple Supervision Sources and Perturbation Experts

Multi-teacher distillation includes instance-dependent teacher selection (Yuan et al., 2021) and learning a distribution to sample one teacher per update (Ding et al., 2024). For on-policy distillation, MOPD and PROMPTSD assign examples to domain- or task-specific teachers (Ma et al., 2026a;b). H-OPD mixes vision-language and text-only teacher distributions (Yin et al., 2026), while Trust-MOPD weights specialist supervision at each student prefix (Sun et al., 2026). Neural Thickets constructs local perturbation experts. RandOpt selects them by task performance and uses majority voting (Gan & Isola, 2026). Its distillation variant uses SFT on generated reasoning traces and answers. NEIGHBORHOOD OPSD selects experts by marginal filtered PTA on reference prefixes and routes one expert's full next-token distribution at each student-visited state.

5 Conclusion

We showed that a privileged teacher's reference-aligned supervision extends beyond its unperturbed parameter setting. NEIGHBORHOOD OPSD turns complementary corrections in this neighborhood into state-wise supervision for a single student. The selected pool achieves higher reference-side selection-feasible coverage than M_0. Student-prefix continuations support using the pool beyond reference trajectories. Matched ablations support the roles of complementary expert selection and state-wise routing in the downstream gains. The highest-peak expert need not provide the best training target, supporting separate choices of anchor direction and support level. Across Qwen3-1.7B, 4B, and 8B, NEIGHBORHOOD OPSD improves the three-benchmark Average@12 over the original OPSD by 2.75, 1.67, and 1.94 points. Inference uses only the distilled student. Appendix F discusses limitations and future work.

AI USE STATEMENT

We used generative AI tools to revise the manuscript, search and summarize related work, create figures, and modify code. We also used them to organize and summarize experimental data and obtain feedback on experiment design and result interpretation. We developed the method and derivations. We reviewed all AI-assisted content and take responsibility for the final text, code, figures, and results.

ETHICS STATEMENT

We study mathematical reasoning with Qwen3 models, OpenThoughts training data, and the AIME and HMMT benchmarks. The experiments involve no human participants.

REPRODUCIBILITY STATEMENT

Section 2 and Appendix A describe the method. Appendix B specifies the data splits, expert construction, training configuration, baseline implementations, and evaluation protocol. Full ablations, parameter sensitivity, and expert analyses appear in Appendices C, D, and E.

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*, 2024.

Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Zixiang Ding, Guoqing Jiang, Shuai Zhang, Lin Guo, and Wei Lin. How to trade off the quantity and capacity of teacher ensemble: Learning categorical distribution to stochastically employ a teacher for distillation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16): 17915–17923, 2024. doi: 10.1609/aaai.v38i16.29746. URL https://doi.org/10.1609/aaai.v38i16.29746.

Yulu Gan and Phillip Isola. Neural thickets: Diverse task experts are dense around pretrained weights. *arXiv preprint arXiv:2603.12228*, 2026.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, et al. OpenThoughts: Data recipes for reasoning models. *arXiv preprint arXiv:2506.04178*, 2025.

Harvard-MIT Mathematics Tournament. HMMT February 2025: Problems and Solutions. https://www.hmmt.org/www/archive/282, 2025.

Qiangqiang He, Jin Li, and MingCai Chen. Calibrating teacher–student discrepancy for on-policy distillation. *arXiv preprint arXiv:2609.21619*, 2026. URL https://arxiv.org/abs/2609.21619.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026. URL https://arxiv.org/abs/2601.20802.

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

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards streamlined distillation for large language models. In *International Conference on Machine Learning*, 2024.

Jongwoo Ko, Tianyi Chen, Sungnyun Kim, Tianyu Ding, Luming Liang, Ilya Zharkov, and Se-Young Yun. DistiLLM-2: A contrastive approach boosts the distillation of LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 31044–31062, 2025. URL https://proceedings.mlr.press/v267/ko25a.html.

Yuhan Li, Mingxu Zhang, Dazhong Shen, and Ying Sun. PHF: Privileged hidden flow for on-policy self-distillation. *arXiv preprint arXiv:2606.29340*, 2026.

Chen Lin, Kedi Chen, and Wei Zhang. ReNIO: Reweighting negative trajectory importance for LLM on-policy distillation. *arXiv preprint arXiv:2606.23104*, 2026.

Xiaogeng Liu, Xinyan Wang, Yingzi Ma, Yechao Zhang, and Chaowei Xiao. When are teacher tokens reliable? position-weighted on-policy self-distillation for reasoning. *arXiv preprint arXiv:2605.21606*, 2026a.

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*, 2026b.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *International Conference on Learning Representations*, 2016.

Wenhan Ma, Jianyu Wei, Liang Zhao, Hailin Zhang, Bangjun Xiao, Lei Li, Qibin Yang, Bofei Gao, Yudong Wang, Rang Li, Jinhao Dong, Zhifang Sui, and Fuli Luo. MOPD: Multi-teacher on-policy distillation for capability integration in LLM post-training. *arXiv preprint arXiv:2606.30406*, 2026a.

Yingzi Ma, Zichen Zhu, Ming Jiang, and Chaowei Xiao. One student, many teachers: Multi-task on-policy distillation via soft-prompt privileged context. *arXiv preprint arXiv:2607.18293*, 2026b.

Mathematical Association of America. 2024 American Invitational Mathematics Examination. https://maa.org/maa-invitational-competitions/, 2024.

Mathematical Association of America. 2025 American Invitational Mathematics Examination. https://maa.org/maa-invitational-competitions/, 2025.

Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2011.

Zhanming Shen, Jintao Tong, Shaotian Yan, Chen Shen, Hao Chen, Wentao Ye, Xiaomeng Hu, Rui Miao, Haobo Wang, Junbo Zhao, Gang Chen, and Jieping Ye. Purified OPSD: On-policy self-distillation without losing how to think. *arXiv preprint arXiv:2607.02234*, 2026.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. *arXiv preprint arXiv:2601.19897*, 2026. URL https://arxiv.org/abs/2601.19897.

Makoto Shing, Kou Misaki, Han Bao, Sho Yokoi, and Takuya Akiba. TAID: Temporally adaptive interpolated distillation for efficient knowledge transfer in language models. In *International Conference on Learning Representations*, 2025.

Jie Sun, Mao Zheng, Mingyang Song, Zeyuan Liu, Gengsheng Li, Houcheng Jiang, Yilin Cheng, Bichuan Feng, Yuchen Cai, Junfeng Fang, and Xiang Wang. Distill what you trust: Reliability-aware multi-teacher on-policy distillation. *arXiv preprint arXiv:2609.23697*, 2026. URL https://arxiv.org/abs/2609.23697.

Kanghui Tian, Siyuan Liu, Tianxiang Jiang, Shuai Dong, Yizhuo Li, Tian Ding, Yuan Guo, Songze Li, Haowen Hou, Congcong Wang, and Yi Wang. What should a self-teacher see? privileged context design for on-policy self-distillation. *arXiv preprint arXiv:2609.25623*, 2026. URL https://arxiv.org/abs/2609.25623.

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

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.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. *arXiv preprint arXiv:2602.12275*, 2026. URL https://arxiv.org/abs/2602.12275.

Qixiang Yin, Huanjin Yao, Yuchen Cai, Jianghao Chen, Ziyi Wang, Min Yang, Fei Su, and Zhicheng Zhao. H-OPD: Confidence aware heterogeneous multi-teacher multimodal on-policy distillation. *arXiv preprint arXiv:2607.02592*, 2026.

Fei Yuan, Linjun Shou, Jian Pei, Wutao Lin, Ming Gong, Yan Fu, and Daxin Jiang. Reinforced multi-teacher selection for knowledge distillation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14284–14291, 2021. doi: 10.1609/aaai.v35i16.17680. URL https://doi.org/10.1609/aaai.v35i16.17680.

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.

A Method Details

A.1 Reference-Side Selection Clip

The OPSD pointwise clip in Eq. 12 acts on each vocabulary-level forward-KL contribution before summation. When $\ell_{t,v} > \kappa$, the clipped value is constant and contributes no gradient. Contributions below the threshold remain unchanged.

The reference-side selection clip gives no credit to candidate-position pairs with $c_{j,t} > \kappa_{\mathrm{sel}}$. It prevents experts from gaining selection credit through sharp reference-token peaks that the clipped online objective may suppress. This is a selection proxy motivated by the OPSD objective. It examines one reference-token contribution under a fixed reference prefix. The online clip instead examines every vocabulary-level contribution from the routed expert at student-visited states. Passing the selection clip does not determine whether an online contribution will be clipped.

A.2 Greedy Marginal Gain

For the selection objective in Eq. 8, the gain from adding expert j to the current pool S is

$$\Delta(j \mid S) = \sum_{(x,z)} \sum_{t} \left[\widetilde{A}_{j,t} - \max_{i \in S} \widetilde{A}_{i,t} \right]_+ . \tag{14}$$

Each step adds the remaining expert with the largest $\Delta(j \mid S)$. We rank all 501 candidates in greedy insertion order and retain the first K. The coverage gains on the analysis set in Appendix D.3 motivate $K = 25$.

A.3 MaxPeak Equivalence

At a student-visited state, each token's maximum expert support is

$$u_t(v) = \max_{j \in S} q_t^j(v), \qquad a_t^* \in \arg\max_{v \in \mathcal{V}} u_t(v). \tag{15}$$

The two maxima can be exchanged:

$$\max_{v \in \mathcal{V}} u_t(v) = \max_{v \in \mathcal{V}} \max_{j \in S} q_t^j(v) = \max_{j \in S} \max_{v \in \mathcal{V}} q_t^j(v). \tag{16}$$

Selecting the highest-peak expert and then its top token therefore gives a maximizer of u_t, as used by MaxPeak in Eq. 9. This online score uses expert probabilities on the current student prefix. Offline selection uses filtered reference-token PTA.

B Complete Experimental Protocol

B.1 Expert Construction and Selection

Selection and audit data. For Qwen3-8B, we randomly sample two disjoint question sets from the OPSD training data. The selection set contains 1,000 questions and the analysis set contains 500. This matches expert selection and analysis to the student training distribution. Their identifiers are `openthoughts_train_selection_1000` ($\mathcal{D}_{\text{sel}}$) and `openthoughts_test_indep_500`, respectively. The analysis set and all three downstream benchmarks are excluded from expert selection.

Candidates and thresholds. For each model scale, we sample 500 perturbation seeds from the scale's own frozen base checkpoint. Together with M_0, the resulting experts form a pool of 501 candidates. Expert selection uses a split disjoint from AIME and HMMT. The 8B radius grid uses $\sigma \in \{.001, .002, .004, .006\}$, $\tau_{\text{sel}} = .99$, and $\kappa_{\text{sel}} = .06$. The first 25 experts in greedy insertion order are retained. The reference-side selection clip excludes candidate-position pairs whose reference-token FKL contribution exceeds the threshold on fixed reference prefixes. The OPSD pointwise clip acts separately on each vocabulary-level forward-KL contribution from the routed expert distribution at student-visited states. The calibrated perturbation radii are $\sigma = .0006$, $.0012$, and $.002$ for Qwen3-1.7B, 4B, and 8B.

B.2 Selector Definitions

Quality Top-K and Greedy sample coverage evaluate one generated response per expert and question. Let $N = |\mathcal{D}_{\text{sel}}|$. For candidate expert j and question i, define

$$\begin{aligned} b_{j,i} &= \mathbb{I}[\text{expert } j \text{ answers question } i \text{ correctly}], \\ Q_j &= \frac{1}{N}\sum_{i=1}^{N} b_{j,i}. \end{aligned} \tag{17}$$

Both selection scores use final-answer correctness without reference-side SCGate or the selection clip.

Quality Top-K. We select the K experts with the highest Q_j. Ties follow the fixed candidate order.

Greedy sample coverage. Let $C_j = \{i : b_{j,i} = 1\}$ be the questions answered correctly by expert j. The selection objective counts questions answered correctly by at least one selected expert:

$$F(S) = \left| \bigcup_{j \in S} C_j \right|. \tag{18}$$

Starting from $S = \varnothing$, each step adds the expert with the largest increase in coverage:

$$j^\star = \arg\max_{j \notin S} \left| C_j \setminus \bigcup_{k \in S} C_k \right|, \qquad S \leftarrow S \cup \{j^\star\}. \tag{19}$$

We recompute the gains after each addition. Ties favor higher Q_j, then the fixed candidate order. Selection continues until K experts are chosen. Once coverage saturates, this tie rule fills the pool by Q_j.

Independent PTA Top-K. We score each expert by its total filtered PTA on the selection set:

$$R_j = \sum_{(x,z) \in \mathcal{D}_{\text{sel}}} \sum_{t} \widetilde{A}_{j,t}. \tag{20}$$

We rank experts once by R_j and select the top K. This rule uses the same reference-side SCGate and selection clip as the PTA selector. Scores remain fixed as experts are added to the pool.

B.3 Token Consensus Routing

Token consensus constructs the training target from the pool S of 25 experts selected by the PTA selector. At each student prefix, every expert casts one vote for its top-1 token:

$$a_t^j = \arg\max_{v \in \mathcal{V}} q_t^j(v), \qquad n_t(v) = \sum_{j \in S} \mathbb{I}[a_t^j = v]. \tag{21}$$

The consensus token is

$$a_t^{\text{cons}} \in \arg\max_{v \in \mathcal{V}} n_t(v). \tag{22}$$

A token can be selected without receiving more than half of the votes. Ties favor the largest $\sum_{j \in S: a_t^j = v} q_t^j(v)$, then the fixed expert order. We collect all experts that support the consensus token:

$$\mathcal{E}_t^{\text{cons}} = \left\{ j \in S : a_t^j = a_t^{\text{cons}} \right\}. \tag{23}$$

We average their full vocabulary distributions with equal weights:

$$\widehat{q}_t^{\text{cons}}(v) = \frac{1}{|\mathcal{E}_t^{\text{cons}}|} \sum_{j \in \mathcal{E}_t^{\text{cons}}} q_t^j(v). \tag{24}$$

As in MaxPeak-anchored quantile routing, the loss denominator is $\sum_t g_t$ (Eq. 13).

B.4 Training Configuration

All three model scales use learning rate 10^{-6} and effective batch size 64. Training lasts 100 optimizer steps with checkpoints saved every 20 steps. The 8B experiments use eight GPUs, per-device batch size 4, and a gradient accumulation factor of 2. For Neighborhood OPSD, we use full-parameter optimization, forward KL ($\beta = 0$ in the generalized divergence implementation), and a full-vocabulary loss. Unless varied, the defaults are $K = 25$, $q = .75$, online SCGate threshold $\tau = .99$, and $\kappa_{\text{sel}} = \kappa = .06$. All expert parameters remain fixed, and only the student is updated. Token-level post-sum clipping is disabled. Gradient checkpointing and bfloat16 are enabled.

Memory and computation. For each expert forward pass, we add its deterministic perturbation, compute the distribution, and restore the base weights. Materializing all K expert distributions would require $O(KBS|\mathcal{V}|)$ memory. We instead store $O(KBS)$ routing metadata on CPU in a first streaming pass and materialize one expert distribution at a time in a second pass. Table 10 reports estimated step time relative to OPSD+SCGate. Appendix D.4 gives the timing measurements and calculation. Inference uses only the distilled student.

Rollouts and filtering. Student rollouts use at most 1024 tokens, temperature 1.1, top-$p = .95$, and top-$k = 20$. Thinking is disabled in the student prompt and enabled in the privileged prompt. The student sees the problem alone, and the teacher receives the same completion with the privileged reference context. In Neighborhood OPSD, SCGate filters a position when the student probability of its sampled token exceeds the threshold. The OPSD+SCGate control applies the same online gate to the fixed M_0 target and has no offline expert-selection stage. In the SCGate ablation, we remove both the reference-side and online gates and reselect the expert pool. In the selection-clip ablation, we reselect the pool without the reference-side selection clip while retaining the online OPSD pointwise clip.

B.5 Baseline Implementations

OPSD, EOPD, and PW-OPSD use the frozen reference-conditioned teacher M_0. The student generates a rollout from the problem alone, and the teacher evaluates that rollout with access to the reference solution and final answer. OPSD applies the clipped forward-KL contribution in Eq. 12 to q_t^0 at every rollout position, with uniform position weights.

For EOPD, we apply the entropy-aware distillation objective of Jin et al. (2026) using this privileged teacher. This comparison tests the objective in the self-distillation setting, with the reference-conditioned M_0 supplying the teacher distributions. PW-OPSD uses the position-dependent loss weighting proposed by Liu et al. (2026a), using the same rollout protocol and reference-conditioned teacher.

B.6 Evaluation and Statistics

Our runs use thinking mode, temperature 1.0, top-$p = .95$, top-$k = -1$, maximum 38,912 new tokens, and `val.n=12`. Following PW-OPSD (Liu et al., 2026a), we report the step-100 checkpoint for NEIGHBORHOOD OPSD, all rerun baselines, and all ablation variants. Base, SFT, and GRPO results are taken from Zhao et al. (2026) under the source paper's evaluation and checkpoint-selection protocols. The other baseline results are our reruns. We conduct three independent training runs for NEIGHBORHOOD OPSD, OPSD, EOPD, and PW-OPSD at each scale. The Qwen3-8B OPSD+SCGate control also uses three independent training runs. We report the mean and sample standard deviation across runs. For Avg., we use each run's three-benchmark mean. The remaining baselines and analysis variants are point estimates unless noted otherwise.

C Complete Ablation Results

The following tables give the results for each benchmark corresponding to the averages in Section 3. All experiments use Qwen3-8B and Average@12. Results are matched single runs unless noted otherwise.

C.1 Expert Selection

All selectors use the same 1,000 selection questions, 501 candidates, $\sigma = .002$, and $K = 25$. Training uses MaxPeak-anchored quantile routing with $q = .75$ and online SCGate with $\tau = .99$.

Table 4: Expert-selection accuracy (%) on Qwen3-8B with Average@12.

Selector	AIME24	AIME25	HMMT25	Avg.
Random Top-K	$77.04 \pm .85$	$70.28 \pm .48$	$45.56 \pm .73$	$64.29 \pm .53$
Quality Top-K	77.50	70.83	46.67	65.00
Greedy sample coverage	77.78	71.67	46.67	65.37
Independent PTA Top-K	77.78	71.39	47.78	65.65
PTA selector (Ours)	**78.33**	**72.78**	**48.61**	**66.57**

Random Top-K: mean $\pm$ sample standard deviation over three pools. Other rows are single runs. Both PTA selectors use the same reference-side SCGate and selection clip.

C.2 Online Routing

We compare Uniform averaging, Random routing, Token consensus, and MaxPeak-anchored quantile routing on the same PTA-selected pool at $K = 25$ and $\sigma = .002$. Random routing draws $\widehat{j}_t$ uniformly from S at each student-visited state and uses the full distribution $\widehat{q}_t = q_t^{\widehat{j}_t}$. It retains the online SCGate and clipped forward-KL objective. The quantile comparison keeps the MaxPeak anchor rule fixed.

C.3 Divergence and Target Distribution

These ablations use the same 100-step training budget. We report the step-100 checkpoint for each run. Both supervision sources use online SCGate with $\tau = .99$. All NEIGHBORHOOD OPSD variants retain the selected pool and router at $\sigma = .002$, $K = 25$, and $q = .75$. Table 6 reports the complete results from single training runs. The full-vocabulary forward-KL reference is the single run reported in Table 7.

Top-N-tail construction. Let q_t denote the teacher target, either q_t^0 for OPSD+SCGate or $\widehat{q}_t$ for NEIGHBORHOOD OPSD. At each state, we select its N most probable tokens, $H_{N,t} = \{v_1, \ldots, v_N\}$. We use this same set to compress the teacher and student distributions into $N + 1$

Table 5: Routing accuracy (%) at $\sigma = .002$ and $K = 25$ on Qwen3-8B.

Routing rule	q	AIME24	AIME25	HMMT25	Avg.
Uniform averaging	n/a	77.78	69.17	46.94	64.63
Random routing	n/a	**78.33**	70.28	46.11	64.91
Token consensus	n/a	77.22	71.39	46.39	65.00
MaxPeak-anchored quantile	0	77.22	70.83	46.11	64.72
	.25	77.50	71.11	46.94	65.19
	.50	78.06	72.22	47.78	66.02
	.75	**78.33**	**72.78**	**48.61**	**66.57**
	1	77.22	71.67	46.94	65.28

All scores use Average@12. MaxPeak selects the anchor token; q selects the rank of its probability gap among anchor-consistent experts.

Table 6: Divergence and target-distribution accuracy (%) on Qwen3-8B.

Source	Setting	AIME24	AIME25	HMMT25	Avg.
OPSD + SCGate	JSD	77.50	70.28	46.94	64.91
N-OPSD	JSD	78.33	71.67	47.50	**65.83**
OPSD + SCGate	RKL	76.94	69.17	44.72	63.61
N-OPSD	RKL	77.22	70.00	45.83	**64.35**
OPSD + SCGate	Top-1-tail	74.44	68.06	46.94	63.15
N-OPSD	Top-1-tail	76.94	71.94	46.94	**65.27**
OPSD + SCGate	Top-100-tail	76.39	72.50	44.17	64.35
N-OPSD	Top-100-tail	77.22	70.83	48.06	**65.37**
N-OPSD	Full-vocabulary FKL	78.33	72.78	48.61	66.57

Average@12; all entries are single runs. JSD and RKL use the full vocabulary; Top-N-tail uses clipped forward KL. The last row repeats the full-vocabulary FKL reference. Avg. is the unweighted mean across benchmarks.

categories:

$$q_{t,i}^{(N)} = q_t(v_i), \quad p_{t,i}^{(N)} = p_t(v_i), \qquad i = 1, \dots, N,$$
$$q_{t,N+1}^{(N)} = \sum_{v \notin H_{N,t}} q_t(v), \quad p_{t,N+1}^{(N)} = \sum_{v \notin H_{N,t}} p_t(v). \tag{25}$$

The final category preserves the total tail probability but does not match the probabilities of individual tokens within the tail. Top-1-tail therefore provides a binary soft target, while Top-100-tail has 101 categories.

Clipping after compression. We compute one forward-KL contribution per category and clip it before summing:

$$\ell_{t,i}^{(N)} = q_{t,i}^{(N)} \log \frac{q_{t,i}^{(N)}}{p_{t,i}^{(N)}}, \qquad L_t^{(N)} = \sum_{i=1}^{N+1} \min(\ell_{t,i}^{(N)}, \kappa), \quad \kappa = .06. \tag{26}$$

The training loss averages $L_t^{(N)}$ over SCGate-retained positions, using the denominator in Eq. 13. The aggregated tail receives one clip, whereas the full-vocabulary objective clips each token separately.

C.4 SCGate and the Reference-Side Selection Clip

OPSD+SCGate applies the online gate to the fixed M_0 teacher. For the component ablations, we reselect the expert pool after removing both SCGate stages or the reference-side selection clip. The online OPSD pointwise clip is retained in both ablations.

Table 7: SCGate control and selection-filter ablations on Qwen3-8B (Average@12, %).

Variant	AIME24	AIME25	HMMT25	Avg.
Three independent runs				
OPSD	77.41 ± .32	71.02 ± .32	45.46 ± .42	64.63 ± .16
OPSD + SCGate	77.87 ± .70	71.11 ± .48	45.74 ± .42	64.91 ± .09
N-OPSD	**78.70 ± .69**	**72.78 ± .28**	**48.24 ± .89**	**66.57 ± .46**
Matched single-run component ablations				
Full N-OPSD	**78.33**	**72.78**	**48.61**	**66.57**
w/o SCGate (both stages)	77.78	71.94	46.11	65.28
w/o selection clip	76.94	70.83	45.00	64.26

First block: mean ± sample standard deviation over three independent runs. Second block: matched single runs at $\sigma = .002$, $K = 25$, $q = .75$. Removing SCGate disables both selection and training gates; removing the reference-side selection clip retains the online OPSD clip.

SCGate improves the fixed-teacher average by 0.28 points. In the matched single-run comparison, removing both SCGate stages reduces the average by approximately 1.3 points. Removing the reference-side selection clip reduces it by 2.31 points.

D Parameter Sensitivity and Training Cost

Downstream accuracy results in this section use Qwen3-8B, Average@12, and matched single runs.

D.1 Perturbation Radius

The perturbation radius controls how much expert behavior changes. Small σ keeps experts close to M_0 and may reveal little complementarity. Larger σ can expose more reference-aligned corrections, but can also increase peak probabilities and expert-student mismatch. We train the full method at each radius while fixing $K = 25$ and $q = .75$. Table 8 tests whether the radius chosen by the student-prefix probe also gives the best downstream result.

Table 8: Perturbation-radius accuracy (%) with $K = 25$ and $q = .75$ (Qwen3-8B, Average@12).

σ	AIME24	AIME25	HMMT25	Avg.
.001	77.78	72.20	47.50	65.83
.002	**78.33**	**72.78**	**48.61**	**66.57**
.004	77.78	70.56	45.28	64.54
.006	76.39	69.72	43.61	63.24

The three-benchmark average reaches 66.57 at $\sigma = .002$, compared with 65.83, 64.54, and 63.24 at .001, .004, and .006. The .002 setting is also best on each benchmark. Both the continuation probe and downstream evaluation favor $\sigma = .002$ among the tested radii.

D.2 SCGate Calibration and Sensitivity

The confidence audit uses the same 331,246 Qwen3-8B reference positions as Appendix E.1. For each threshold, we select positions where the student's top prediction reaches that confidence and measure top-1 accuracy on those positions. The downstream columns use the same threshold for online SCGate.

At $\tau = .99$, 217,400 reference positions (65.63%) are in the high-confidence group. The problem-only model reaches 97.67% top-1 accuracy in this group. This motivates focusing supervision on lower-confidence positions.

Table 9: SCGate confidence audit and downstream accuracy at $\sigma = .002$, $K = 25$, $q = .75$.

τ	High-confidence reference positions			Downstream Average@12 (%)			
	Count	Share (%)	Acc. (%)	AIME24	AIME25	HMMT25	Avg.
.95	242,255	73.13	96.02	77.78	71.39	46.67	65.28
.98	227,137	68.57	97.09	77.50	72.22	47.50	65.74
.99	217,400	65.63	97.67	78.33	72.78	48.61	66.57
.995	207,988	62.79	98.10	77.78	71.11	47.22	65.37

D.3 Expert Count

Table 10 reports selection-feasible coverage, downstream accuracy, and estimated step time as the expert pool grows. We measure coverage on the 500-question analysis set using the first K experts in the greedy order determined on the selection set. Coverage is normalized by $K = 50$ and estimated step time by OPSD+SCGate.

Table 10: Expert-count ablation at $\sigma = .002$ on Qwen3-8B (Average@12, %).

K	q	Rel. SF cov. (%)	Est. rel. step time	AIME24	AIME25	HMMT25	Avg.
1	.75	63.47	1.010×	76.94	70.28	46.39	64.54
5	.75	76.63	1.082×	77.78	70.83	47.22	65.28
10	.75	84.48	1.193×	77.50	71.67	47.78	65.65
25	.75	94.81	1.423×	78.33	**72.78**	48.61	**66.57**
	.75			77.78	70.83	44.44	64.35
50	.50	100.00	1.914×	**78.89**	69.44	**49.17**	65.83
	.25			77.50	71.94	45.83	65.09

Coverage is normalized by $K = 50$ and step time by OPSD+SCGate. Timing uses the same student trajectories (Appendix D.4). Lower q values at $K = 50$ probe the relation between quantile and pool size (Section 3.4).

At $K = 25$, the pool retains 94.81% of the $K = 50$ selection-feasible coverage. It reaches the highest tested Average@12 of 66.57, compared with 65.83 for the best tested $K = 50$ setting. Its

estimated step time is $1.423\times$ that of OPSD+SCGate and 25.63% lower than at $K = 50$. The best tested quantile changes from $q = .75$ at $K = 25$ to $q = .50$ at $K = 50$. At $K = 50$, $q = .50$ improves Average@12 by 1.48 points over $q = .75$.

Table 11: Coverage gains on the 500-question analysis set at $\sigma = .002$.

Expert rank K	1	5	10	25	50
Marginal share (%)	63.47	2.18	1.21	.39	.15
Cumulative share (%)	63.47	76.63	84.48	94.81	100.00

Marginal share measures the coverage added by the expert at that rank. Both shares are normalized by cumulative selection-feasible coverage at $K = 50$.

The marginal share falls from 63.47% for the first expert to $.39\%$ at rank 25 and $.15\%$ at rank 50. The curve shows that most available selection-feasible coverage is already present at $K = 25$.

D.4 Training-Step Timing

We measure teacher time on the same cached student trajectories. The unperturbed teacher M_0 used by OPSD+SCGate takes 2.0184 seconds in one forward pass. For Neighborhood OPSD, Pass 1 computes routing metadata, and Pass 2 reruns experts to build the target distribution. Their sum is the teacher time in Table 12.

We estimate step time using the same rollout time (78.215 seconds) and student computation time (82.443 seconds) for all settings. Student computation includes the forward pass, loss computation, backward pass, and other training operations. Adding the measured M_0 teacher time gives an OPSD+SCGate baseline of 162.6764 seconds. For Neighborhood OPSD, we add 0.012 seconds for routing, so estimated step time equals teacher time plus 160.670 seconds. Both teacher time and estimated step time are normalized by their OPSD+SCGate values.

Table 12: Measured teacher time and estimated training-step time on Qwen3-8B.

Setting	Pass 1 (s)	Pass 2 (s)	Teacher time (s)	Relative teacher time	Est. step time (s)	Est. relative step time
M_0	n/a	n/a	2.0184	1.0000×	162.6764	1.000×
$K = 1$	1.754	1.850	3.604	1.7856×	164.274	1.010×
$K = 5$	7.840	7.526	15.366	7.6130×	176.036	1.082×
$K = 10$	16.732	16.688	33.420	16.5577×	194.090	1.193×
$K = 25$	35.221	35.637	70.858	35.1060×	231.528	1.423×
$K = 50$	76.385	74.277	150.662	74.6443×	311.332	1.914×

M_0 is the OPSD+SCGate baseline; each K row uses Neighborhood OPSD. Teacher time is Pass 1 plus Pass 2. Both time ratios are normalized by OPSD+SCGate.

Table 13 gives the estimated step time breakdown at $K = 50$.

Table 13: Estimated training-step breakdown on Qwen3-8B at $K = 50$.

Component	Time (s)	Share (%)
Rollout	78.215	25.1227
Teacher time (Pass 1 + Pass 2)	150.662	48.3927
Routing	0.012	0.0039
Student computation	82.443	26.4807
Total	311.332	100.0000

E Additional Expert Analyses

E.1 Full Reference-Side Audit

We audit the selected expert pools on the 500-question analysis set described in Appendix B. Its complete reference solutions contain 331,246 token positions before filtering.

Table 14: Reference-side audit of the unperturbed teacher and selected expert pools.

σ	SF coverage (%)		Token accuracy (%)		JS(M_j, M_0)	MaxPeak peak
	M_0	Pool	M_0 top-1	MaxPeak anchor		
.001	15.69	21.86	91.33	94.20	.00111	.9762
.002	15.69	26.53	91.33	95.58	.00393	.9871
.004	15.69	29.02	91.33	96.53	.01198	.9962
.006	15.69	31.39	91.33	97.31	.03474	.9993

Selection-feasible (SF) coverage requires reference-side SCGate retention, positive PTA, and the selection clip. Accuracy, JS divergence, and peak probability use all SCGate-retained positions. M_0 is the unperturbed privileged teacher; Pool uses the selected perturbation experts.

At every tested radius, the selected pool has higher selection-feasible coverage than M_0, and MaxPeak has higher anchor accuracy. Larger radii also increase JS divergence and MaxPeak peak probability.

E.2 Complete Raw MaxPeak Radius Probe

Prefix construction. We generate trajectories on the NEIGHBORHOOD OPSD training data using the frozen Qwen3-8B base student. All trajectories use the same generation protocol. After filtering and deduplication, we randomly sample 150 trajectories from each of the following groups. The 300 trajectories come from 300 distinct questions.

- `complete_wrong`: The trajectory contains `<|im_end|>`, and its last boxed answer differs from the reference answer. We retain the first 75% of its tokens as the continuation prefix.
- `truncated_unknown`: The trajectory has no `<|im_end|>`, usually because it reaches the 1,024-token generation limit. Its final correctness is unknown. We retain its complete token sequence as the continuation prefix.

Continuation evaluation. We evaluate each radius on the same 300 prefixes. Raw MaxPeak routes the highest-peak expert at each decoding step and does not use the full router's quantile selection. Both M_0 and Raw MaxPeak receive the same privileged reference solution and may generate up to 4096 new tokens. This budget includes reasoning tokens and excludes the student prefix and other input prompts. A continuation is correct when its first closed boxed answer matches the reference answer. Unfinished generations remain in the denominator. Completion means producing a closed boxed answer within this budget. The FKL proxy uses the union of retained top probabilities and one residual-mass bucket.

Table 15: Raw MaxPeak radius probe on 300 student-prefix continuations.

σ	FKL proxy			Accuracy (%)		Completion (%)	
	M_0	Raw MP	Ratio	M_0	Raw MP	M_0	Raw MP
.001	.18138	.18528	1.02×	83.33	81.33	90.00	88.67
.002	.18138	.20918	1.15×	83.33	**86.67**	90.00	**94.67**
.004	.18138	.29848	1.65×	83.33	62.33	90.00	79.00
.006	.18138	.66341	3.66×	83.33	27.33	90.00	39.33

Raw MP denotes Raw MaxPeak; the proxy ratio is Raw MP/M_0. All continuations are included.

Raw MaxPeak gives the best continuation results at $\sigma = .002$. Compared with M_0, accuracy is 3.33 points higher and completion is 4.67 points higher. At larger radii, the FKL proxy rises and continuation quality falls, even though reference-side anchor accuracy continues to improve.

E.3 Frequent Newly Covered Reference Tokens

Table 16 lists frequent reference tokens at positions newly covered by the perturbation pool at $\sigma = .002$. They include function words such as `the` and `we`, LaTeX delimiters, numerals, punctuation, and reasoning verbs such as `find` and `need`. These are standard lexical and symbolic components of mathematical solutions. The new coverage is therefore not limited to self-reflective style markers such as `hmm` and `wait` (Zhao et al., 2026).

Table 16: Frequent newly covered reference tokens at $\sigma = .002$.

Token	Approx. count	Share (%)	Token	Approx. count	Share (%)
`the`	406	4.70	`)`	107	1.24
`\`	346	4.01	`2`	92	1.06
`we`	339	3.93	`and`	76	.88
`,`	158	1.83	`1`	69	.80
`\(`	134	1.55	`find`	54	.62
`The`	131	1.52	`need`	49	.57
`We`	116	1.34			

Shares use an 8,638-token subset; approximate counts are derived from the shares and rounded to the nearest integer. Displayed tokens omit whitespace.

F Limitations and Future Work

Our experiments cover mathematical reasoning tasks and the Qwen3 model family. Evaluating NEIGHBORHOOD OPSD on other model families and tasks, including code generation, multi-hop question answering, and tool use, would test its broader applicability.

Offline expert selection computes filtered PTA relative to the fixed problem-only distribution p_t^S on reference prefixes. Online routing evaluates experts at states visited by the current student, but the selected expert pool remains fixed throughout training. As the student changes, the experts most useful to it may differ from those selected offline. Periodically recomputing PTA and reselecting experts could reduce this mismatch. However, repeated evaluation of all candidates would increase computational cost. Future work can study efficient pool updates that balance this cost with gains in student performance.