On-policy Distillation

李泊桥

2026.5.15

Why we need OPD

- Post-train method:
  - SFT / Off-policy Distillation: dense supervision, imitate instead of explore & exploit
  - RLVR: sparse reward / credit assignment

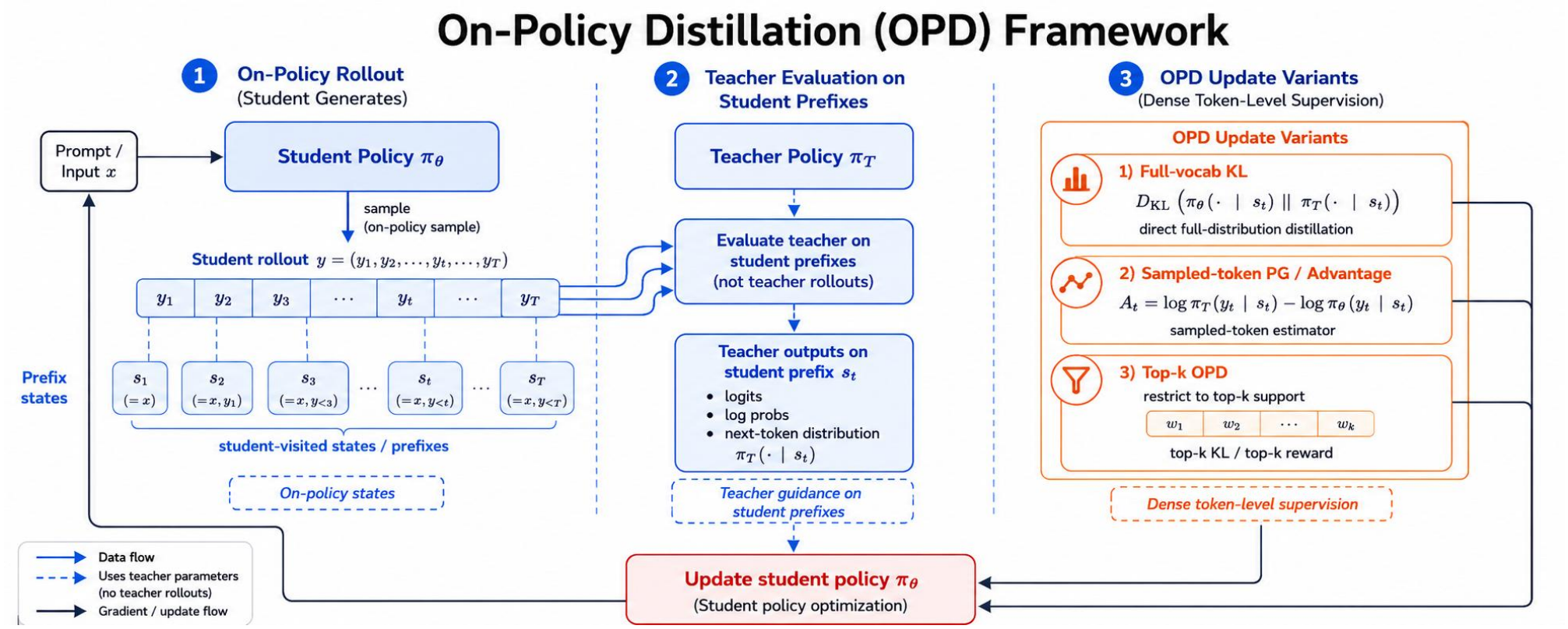
- On-policy Distillation:
  - **Token-level dense** signal (student vs teacher), no need for the manually-designed reward.
  - On-policy: student-generated trajectory

Method	**Sampling**	**Reward signal**
SFT	Off-policy	**Dense**
RL	**On-policy**	Sparse
OPD	**On-policy**	**Dense**

Brief Introduction of OPD

- To be concise:

  On-policy Distillation = student samples, teacher judges



generated by gpt-image-2

Definition & Notations

$\pi_\theta(\cdot \mid s_t)$: student policy, $\pi_T(\cdot \mid s_t)$: teacher policy, $s_t = (x, y_{<t})$

Student sampled response: $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_T) \sim \pi_\theta(\cdot \mid x)$

Token Distribution: $p_t(v) \triangleq \pi_\theta(v \mid x, \hat{y}_{<t})$ $q_t(v) \triangleq \pi_T(v \mid x, \hat{y}_{<t})$

Token-level objective: $\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{x \sim \mathcal{D}_x,\, \hat{y} \sim \pi_\theta(\cdot|x)} \left[\sum_{t=1}^{T} D_{\mathrm{KL}}(p_t \| q_t) \right]$

Optimization

$$\mathcal{L}_t^{\mathrm{OPD}} = D_{\mathrm{KL}}(p_t \| q_t) = \mathbb{E}_{v \sim p_t}\left[\log p_t(v) - \log q_t(v)\right]$$

common objective

- Full-vocab KL (low-variance but expensive)：

$$\mathcal{L}_t^{\mathrm{full}} = \sum_{v \in \mathcal{V}} p_t(v)[\log p_t(v) - \log q_t(v)]$$

- Sampled-token Advantege (cheap but noisy)：

$$A_t^{\mathrm{OPD}} = \mathrm{sg}[\log q_t(\hat{y}_t) - \log p_t(\hat{y}_t)]$$

$$\nabla_\theta \log \pi_\theta(\hat{y}_t | s_t) \cdot A_t^{\mathrm{OPD}}$$

- Balanced top-k:

$$K_t = \mathrm{TopK}_k(p_t)$$

$$\tilde{p}_t(v) = \frac{p_t(v)}{\sum_{u \in K_t} p_t(u)}, \qquad \tilde{q}_t(v) = \frac{q_t(v)}{\sum_{u \in K_t} q_t(u)}$$

$$\mathcal{L}_t^{\text{top-}k} = D_{\mathrm{KL}}(\tilde{p}_t \| \tilde{q}_t)$$

OPD in foundation-model post-training

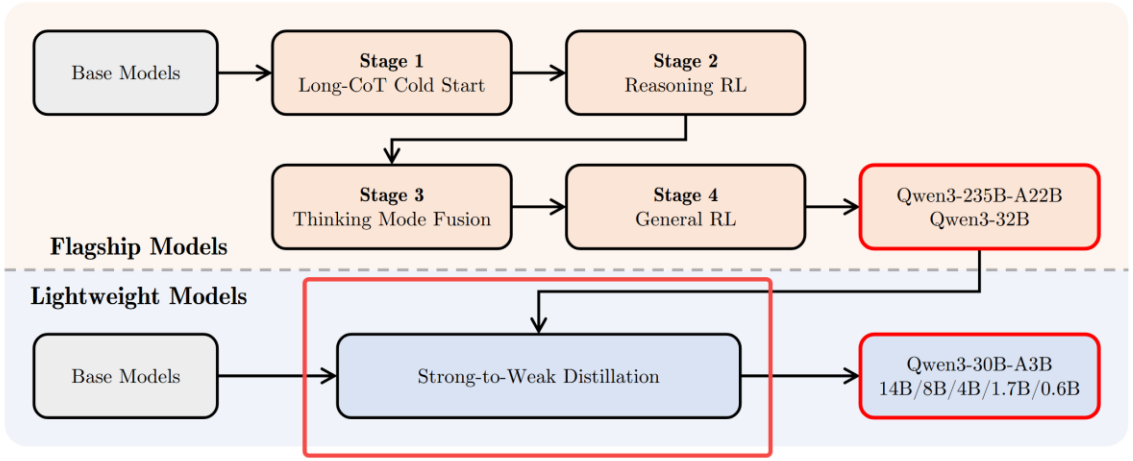


Figure 1: Post-training pipeline of the Qwen3 series models.

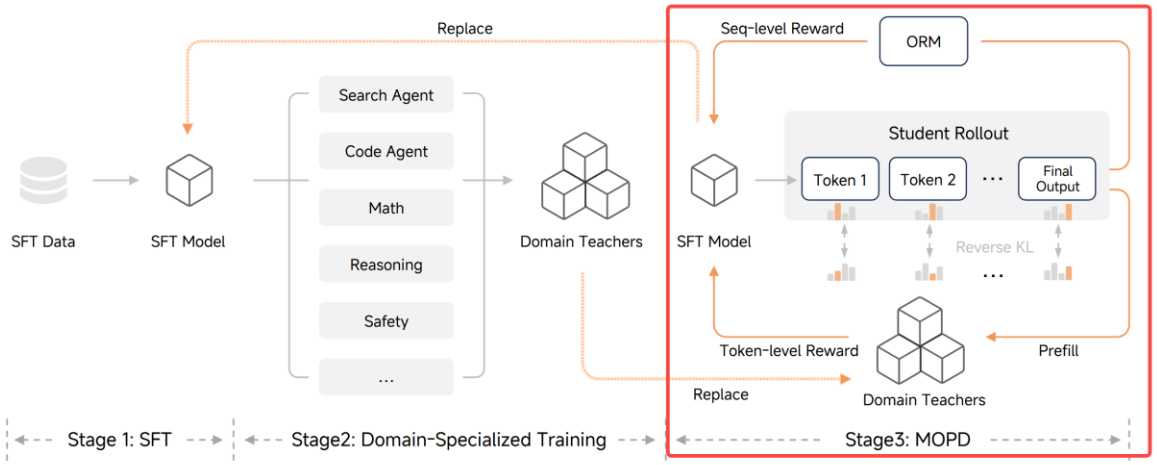


Figure 3 Overview of MiMo-V2-Flash post-training stages.

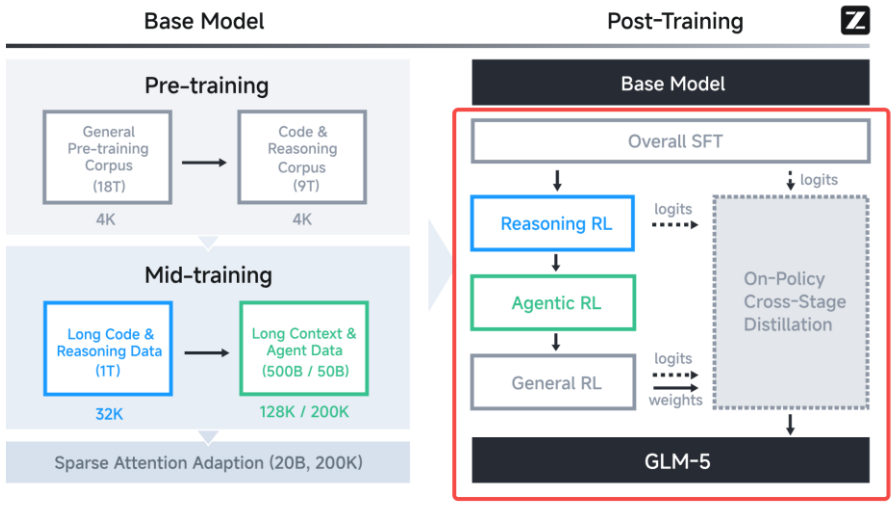


Figure 5: Overall training pipeline of GLM-5.

5.1. Post-Training Pipeline

Following pre-training, we conducted a post-training phase to yield the final models of DeepSeek-V4 series. Although the training pipeline largely mirrored that of DeepSeek-V3.2, a critical methodological substitution was made: the mixed Reinforcement Learning (RL) stage was entirely replaced by On-Policy Distillation (OPD).

DeepSeek-V4

4.1 Understanding Training

To achieve a seamless integration of understanding and generation, we optimize the unified understanding–generation paradigm within the MLLM framework through a multi-stage training strategy. Initialized from the Qwen-VL (Bai et al., 2025b) base model (prior to post-training), our approach comprises three phases: CT to bolster multi-modal understanding, multi-task SFT to generalize generation-oriented tasks, and on-policy multi-teacher distillation to refine real-world performance.

Wan-image

OPD: Effective but Still Poorly Understood

- DeepSeek V4: Full-vocab reverse KL to stable training
- GLM-5: Sampled token, optimize through policy gradient
- MiMo-V2: Sampled token + ORM
- Qwen 3: Off-policy distillation + On-policy distillation

However, many attempts to reproduce OPD experiments or adapt OPD to other tasks failed.

When can OPD succeed, and why?

Rethinking On-Policy Distillation of Large Language Models: Phenomenology, Mechanism, and Recipe

Yaxuan Li[*1,2], **Yuxin Zuo**[*†1], **Bingxiang He**[*†1], **Jinqian Zhang**[1], **Chaojun Xiao**[‡1], **Cheng Qian**[3], **Tianyu Yu**[1], **Huan-ang Gao**[1], **Wenkai Yang**[4], **Zhiyuan Liu**[‡1], **Ning Ding**[‡1]

[1]Tsinghua University [2]ShanghaiTech University [3]University of Illinois Urbana-Champaign [4]Renmin University of China

[*]**Equal Contribution.** [†]**Project Lead.** [‡]**Corresponding Authors.**

Code: https://github.com/thunlp/OPD.

hebx24@mails.tsinghua.edu.cn, {xcj,liuzy,dingning}@tsinghua.edu.cn

arxiv.org/abs/2604.13016

Motivation: Revealing OPD Success and Failure via Training Dynamics

OPD assumes that teacher signals on student-visited states are useful.

But several questions remain open:

- Does better performance always imply a better teacher?
- What kind of teacher can provide high-quality supervision?
- What is actually learned during OPD training?
- How to make OPD successful?

Rethinking OPD answers these questions by analyzing teacher-student training dynamics.

Dynamic Metrics (gap between teacher and student)

Overlap Ratio

$$\mathcal{M}_{\text{overlap}} \triangleq \mathbb{E}_t \left[\frac{|S_t^{(p)} \cap S_t^{(q)}|}{k} \right].$$

Average proportion of tokens appear simultaneously in both the student's and the teacher's top-k sets.

Overlap-token advantage

$$\mathcal{M}_{\text{adv}} \triangleq \mathbb{E}_t \left[\frac{1}{|S_t^{(p)} \cap S_t^{(q)}|} \sum_{v \in S_t^{(p)} \cap S_t^{(q)}} A_t(v) \right]$$

$$A_t(v) \triangleq \bar{p}_t(v)(\log \bar{q}_t(v) - \log \bar{p}_t(v))$$

Where $\bar{p}_t, \bar{q}_t$ are renormalized stu&tea distributions over the intersection tokens

Entropy Gap

$$\Delta H_t = |H(q_t) - H(p_t)|$$

Two conditions make OPD successful

- **Thinking-pattern consistency**

  The student and teacher should share compatible thinking patterns.

- **Teacher with new knowledge**

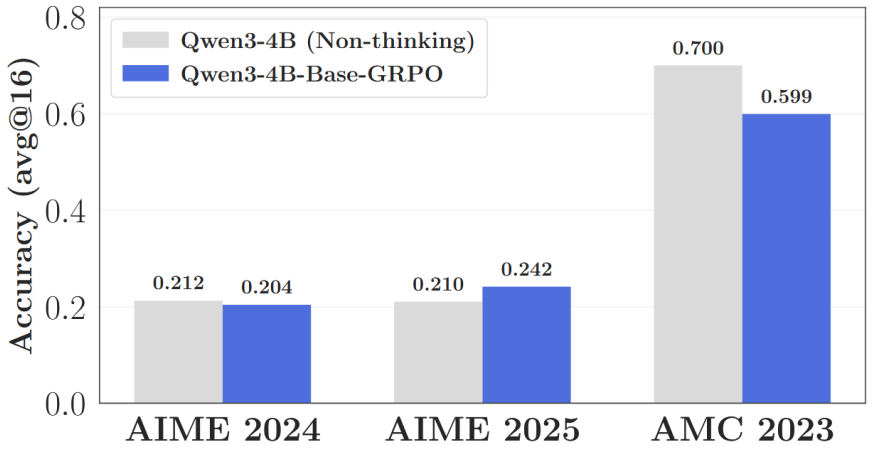
  Teacher should contain something new to transfer.

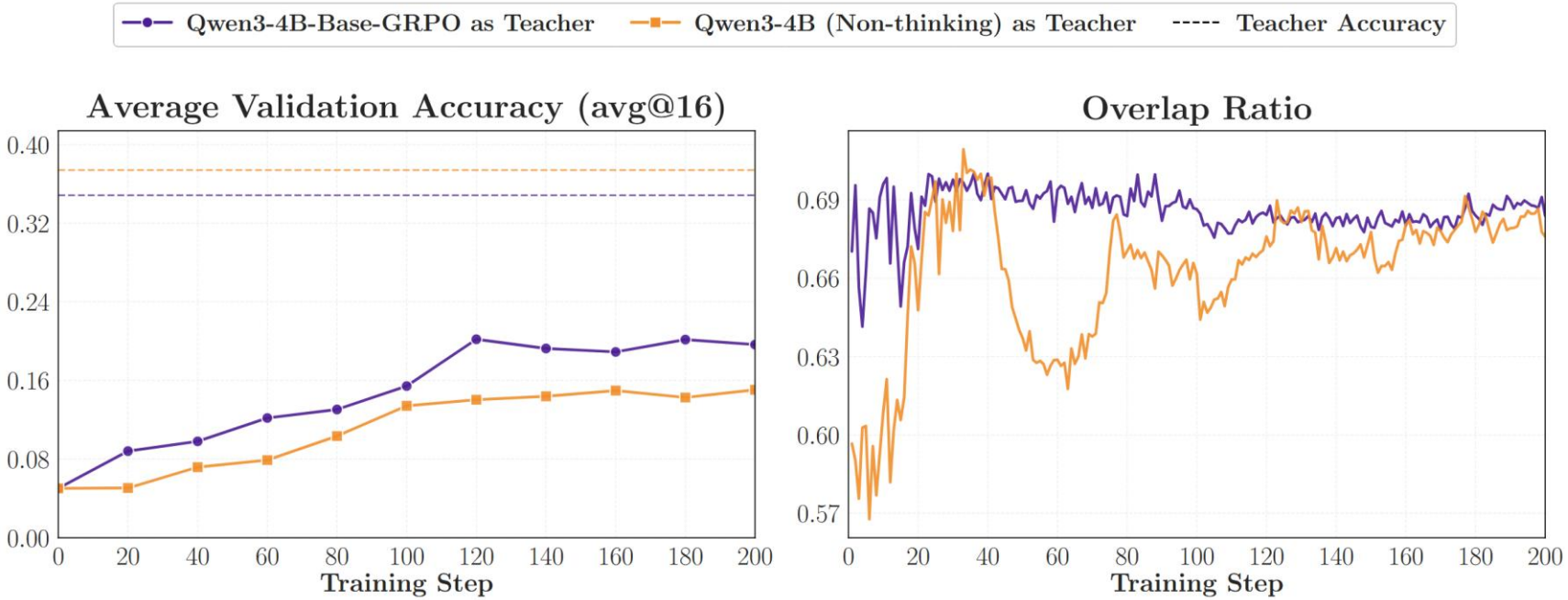
Thinking-pattern consistency

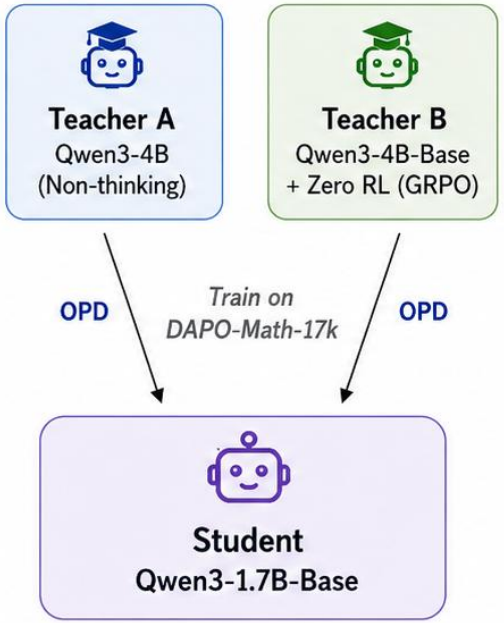
Same-size teacher model with comparable performance on math tasks.

4B-Base-GRPO and 1.7B share similar thinking pattern (higher initial overlap ratio)

The result suggesting that **early-stage thinking-pattern mismatch causes a loss of distillation benefit that cannot be recovered later.**







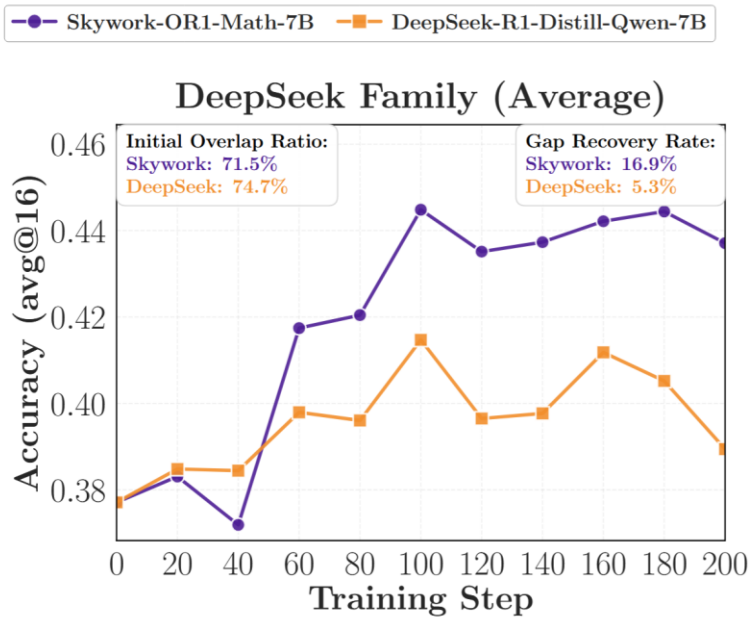
New Knowledge, Not Just Scale

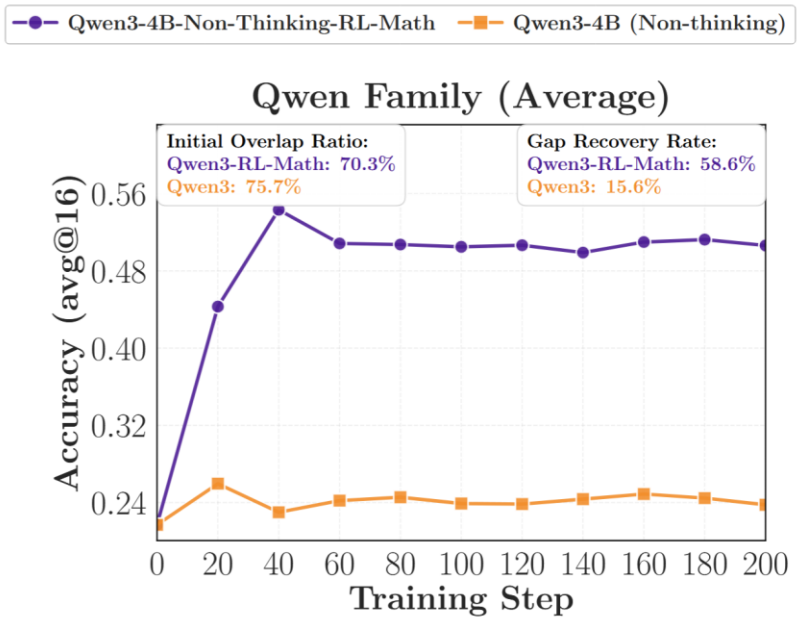
Teacher1 = bigger size student (compatible thinking pattern)

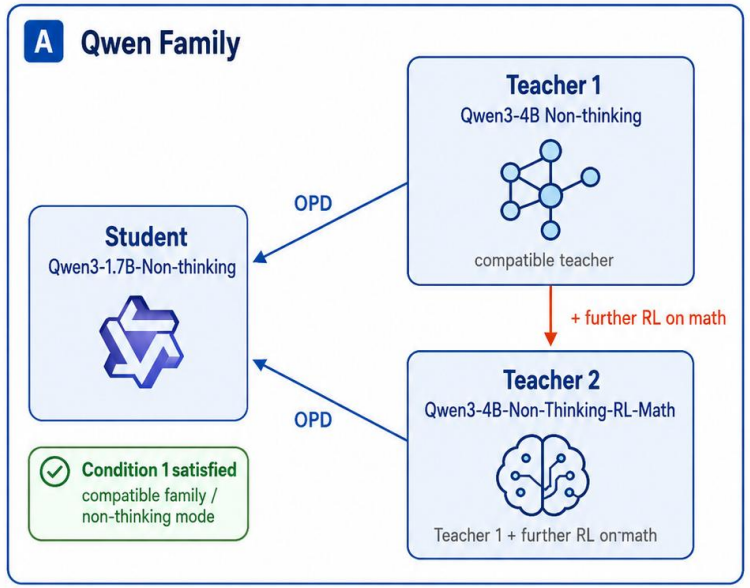
Teacher2 = Teacher1 + RL (new knowledge)

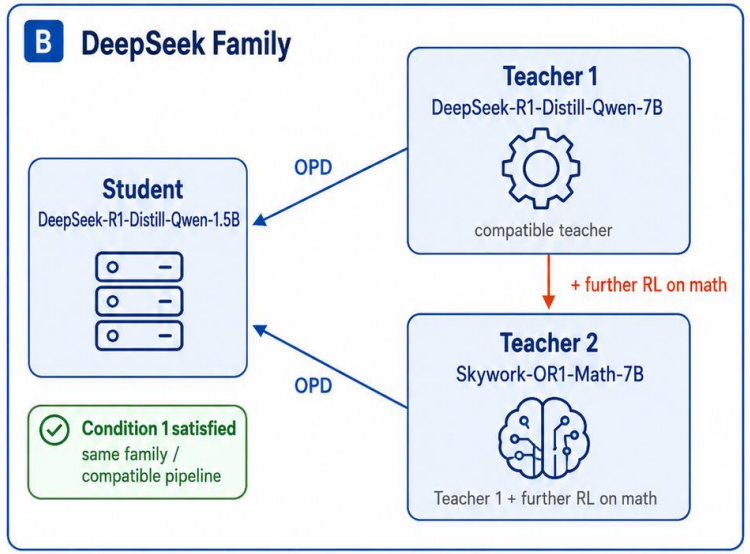
If compatibility were sufficient, Teacher 1 should already work well.

The stronger gains from Teacher 2 suggest that **additional RL-acquired capability provides the missing transferable signal.**





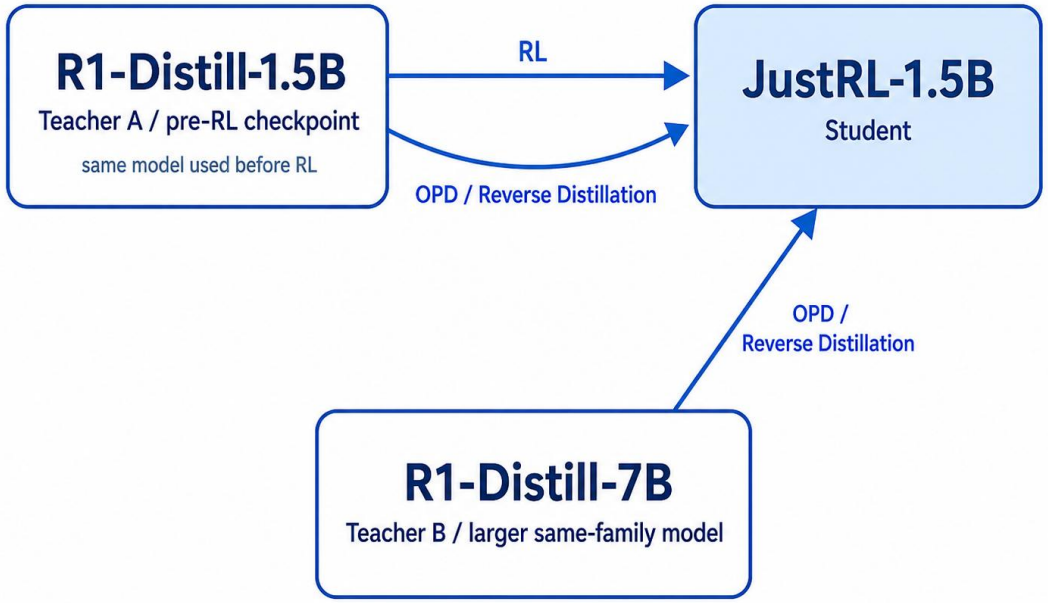


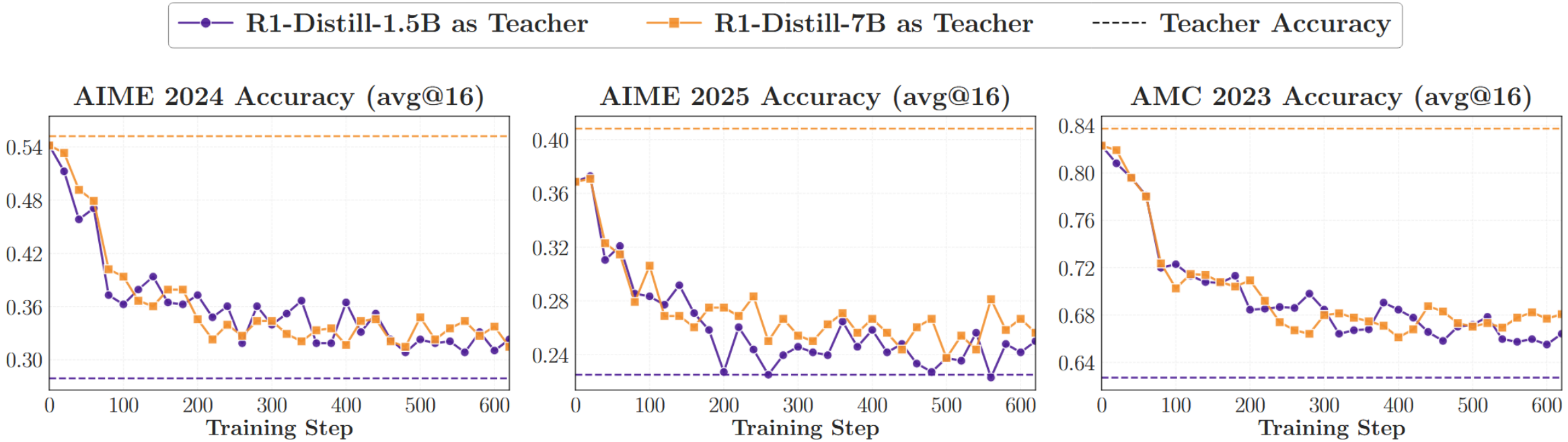


$$Recovery\ rate = \frac{Acc_{after\ OPD} - Acc_{before\ OPD}}{Acc_{teacher} - Acc_{before\ OPD}}$$

Validation: reverse distillation

- OPD fundamentally learns thinking patterns.
- Benchmark performance does not predict OPD outcome.
- Higher scores do not imply new knowledge for OPD.





Mechanism: What Does OPD Actually Learn?

- **Progressive alignment**

  Successful OPD is accompanied by progressive alignment of **high-probability tokens** on student-visited prefixes.

$$\text{overlap ratio} \uparrow \quad \text{overlap-token advantage} \to 0 \quad \text{Entropy Gap} \downarrow$$

- **Overlap sufficiency**

  The main optimization effect concentrates on the **shared top-k tokens.**

$$\text{Optimize overlap tokens only} \approx \text{standard Student Top-k OPD}$$

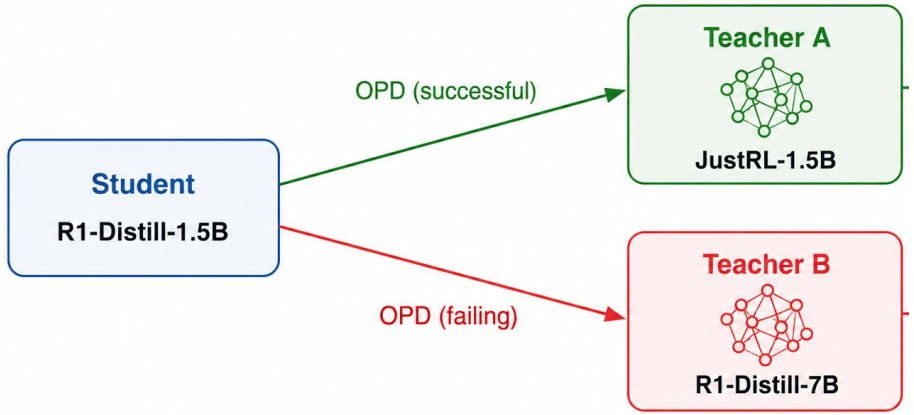
$$\text{Optimize non-overlap tokens} \ll \text{standard OPD}$$

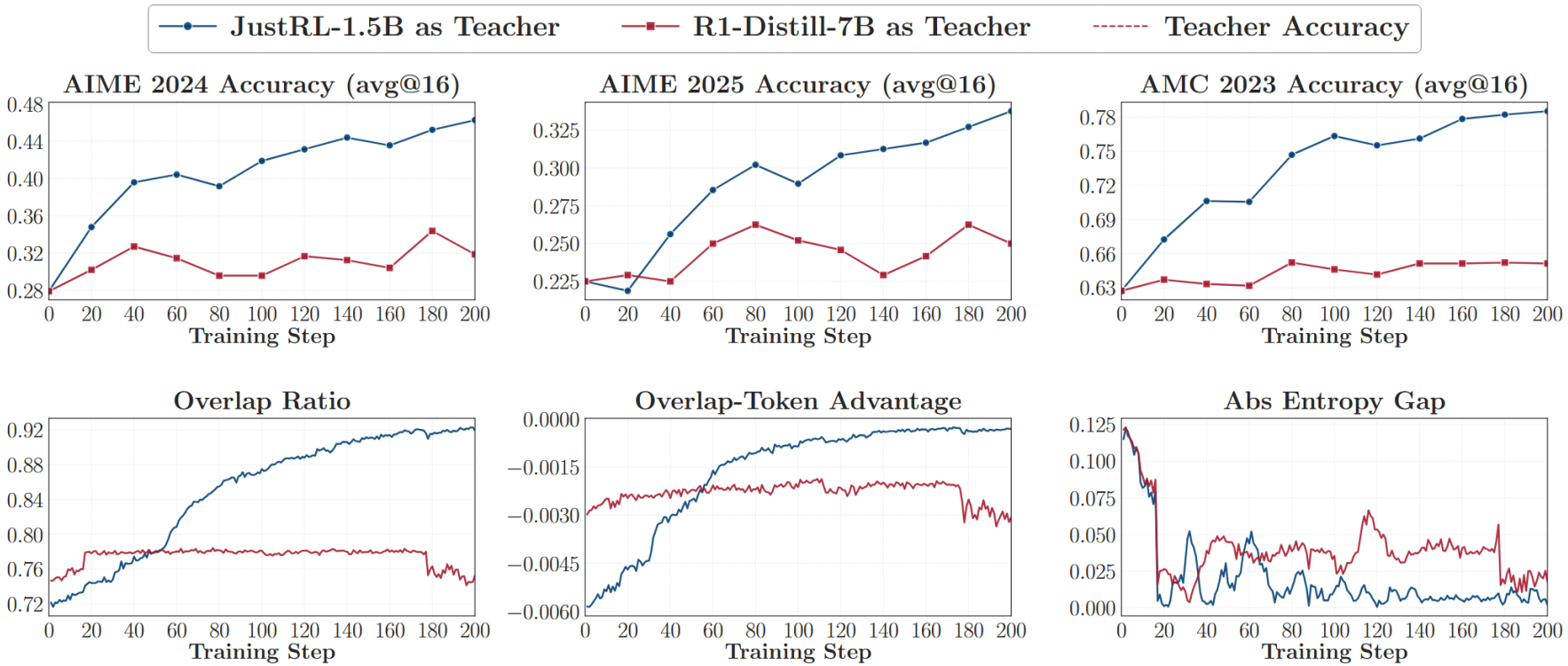
Dynamics of OPD

Successful OPD:

- overlap ratio↑
- overlap-token advantage → 0
- Entropy Gap ↓

Progressive Alignment of High-Probability Tokens





Overlap-token advantage

$$\mathcal{M}_{\text{adv}} \triangleq \mathbb{E}_t\left[\frac{1}{|S_t^{(p)} \cap S_t^{(q)}|} \sum_{v \in S_t^{(p)} \cap S_t^{(q)}} A_t(v)\right]$$

$$A_t(v) \triangleq \bar{p}_t(v)(\log \bar{q}_t(v) - \log \bar{p}_t(v))$$

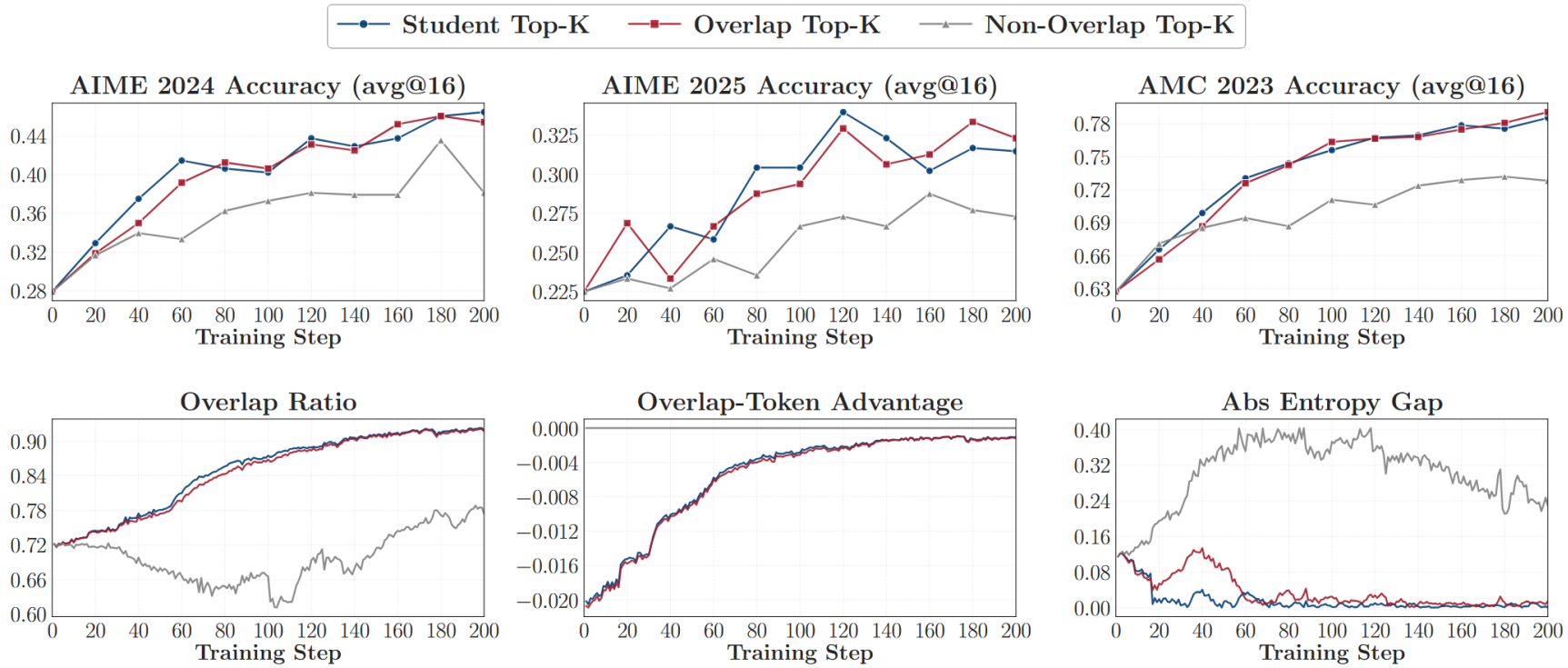
Where $\bar{p}_t, \bar{q}_t$ are renormalized stu&tea distributions over the intersection tokens

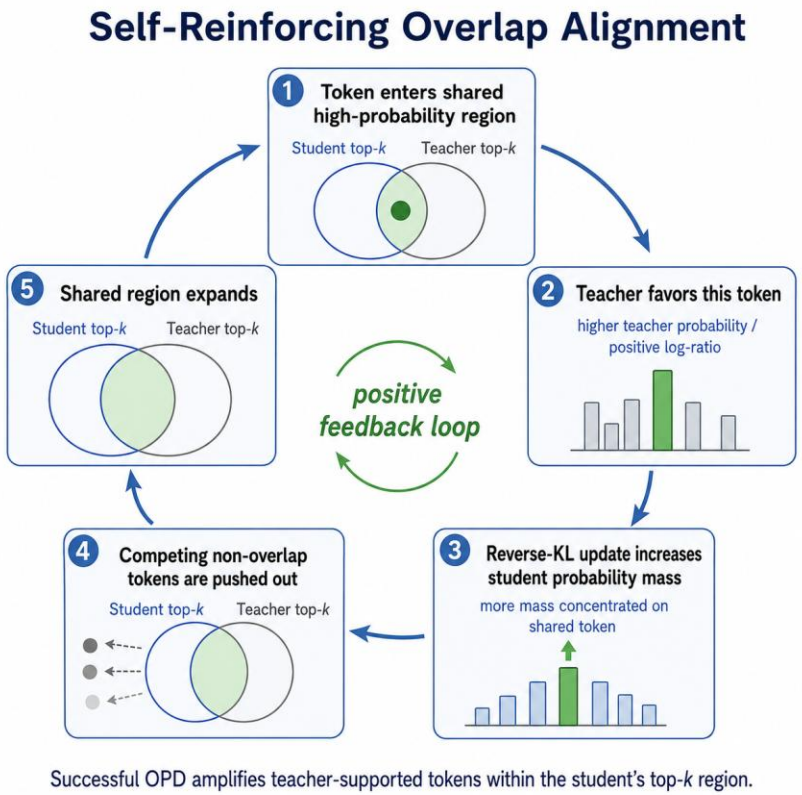
Optimizing Shared Tokens Alone Suffices

Distillation Loss only covers:

- Student Top-k
- Overlap Top-k
- Non-Overlap Top-l

Overlap tokens are the main region where OPD actually works.





Successful OPD amplifies teacher-supported tokens within the student's top-*k* region.

How to improve OPD training

- Off-policy cold start
  To align thinking pattern

- Teacher-aligned prompts
  To provide better supervision

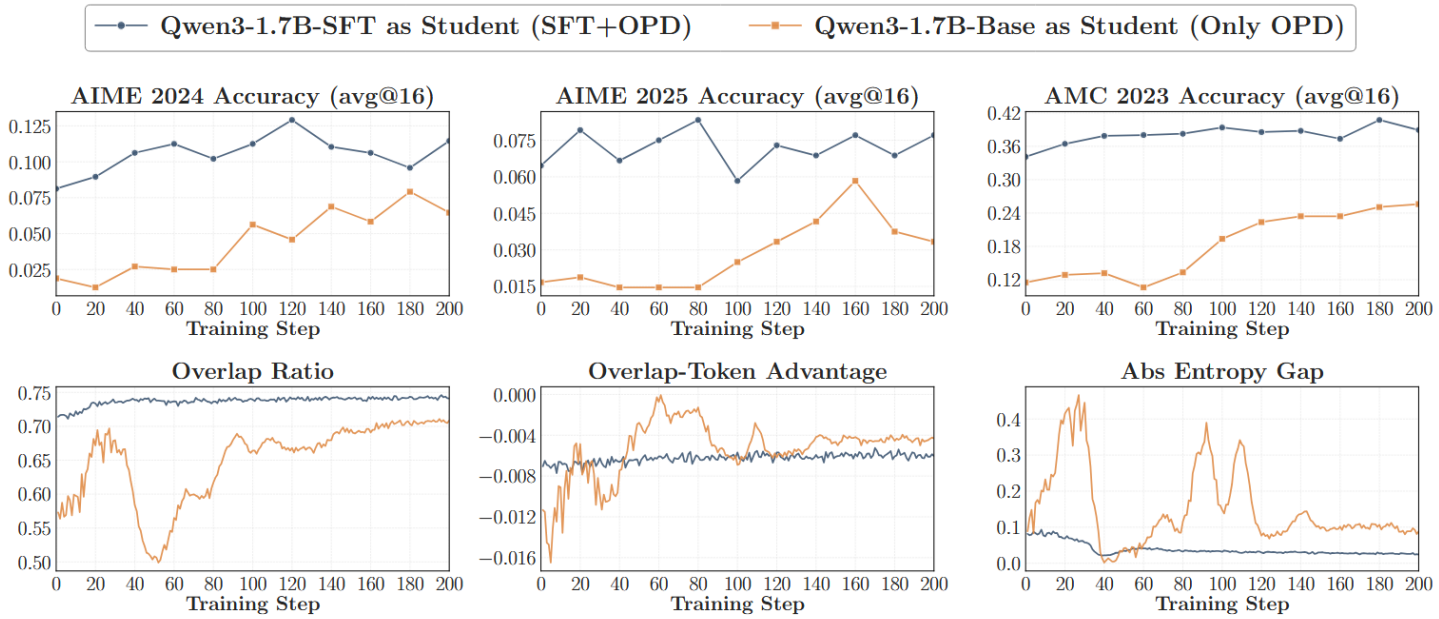


Figure 8 | Effect of off-policy cold start before OPD, using fixed teacher Qwen3-4B (Non-thinking). The two curves correspond to OPD from Qwen3-1.7B-SFT and Qwen3-1.7B-Base.

Limitation & Discussion

1. Reward Quality Degrades with Trajectory Depth

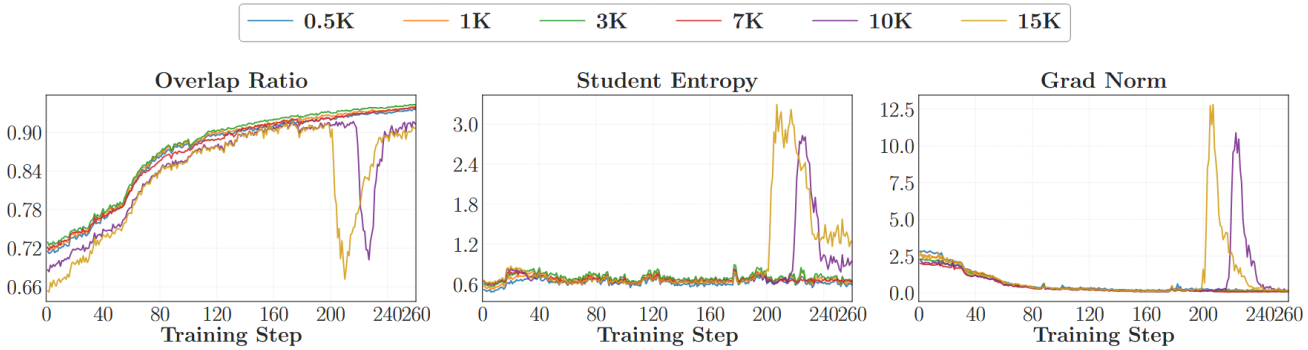
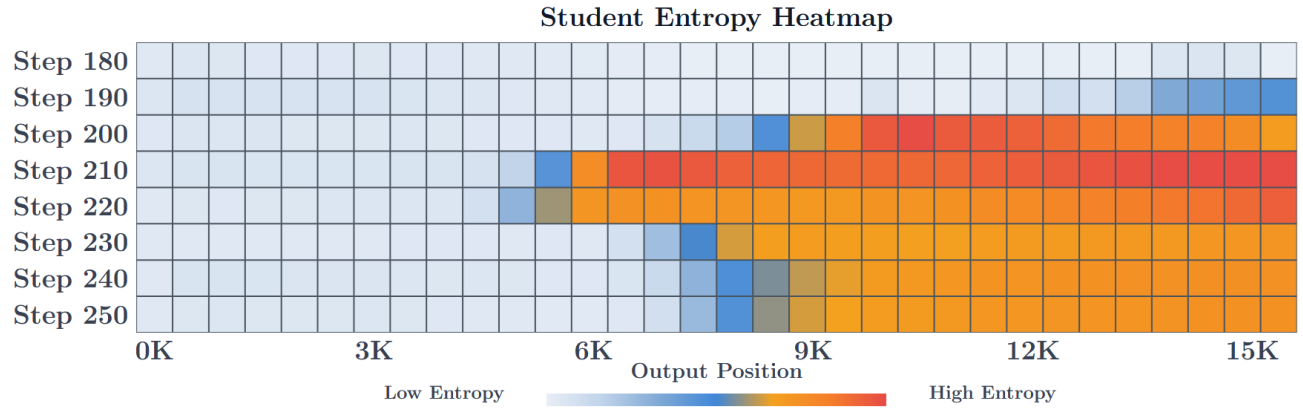
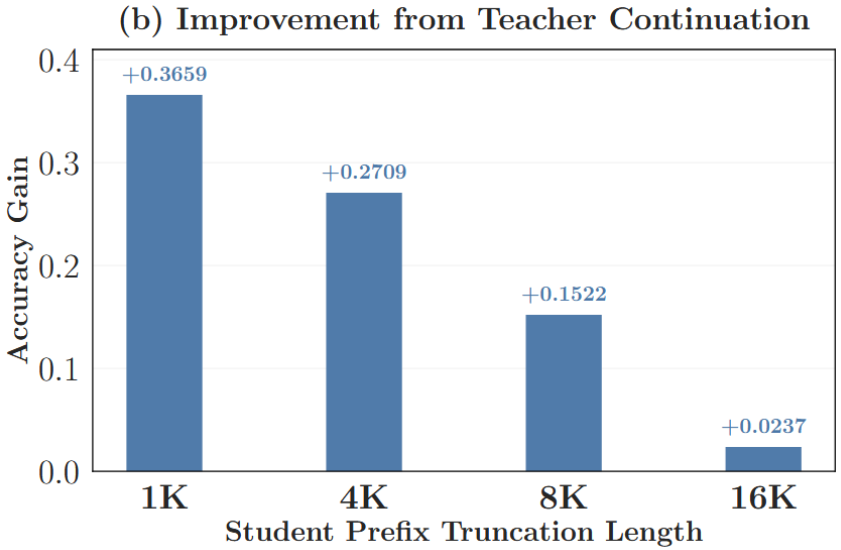


Figure 12 | Training dynamics under different maximum response lengths for OPD.



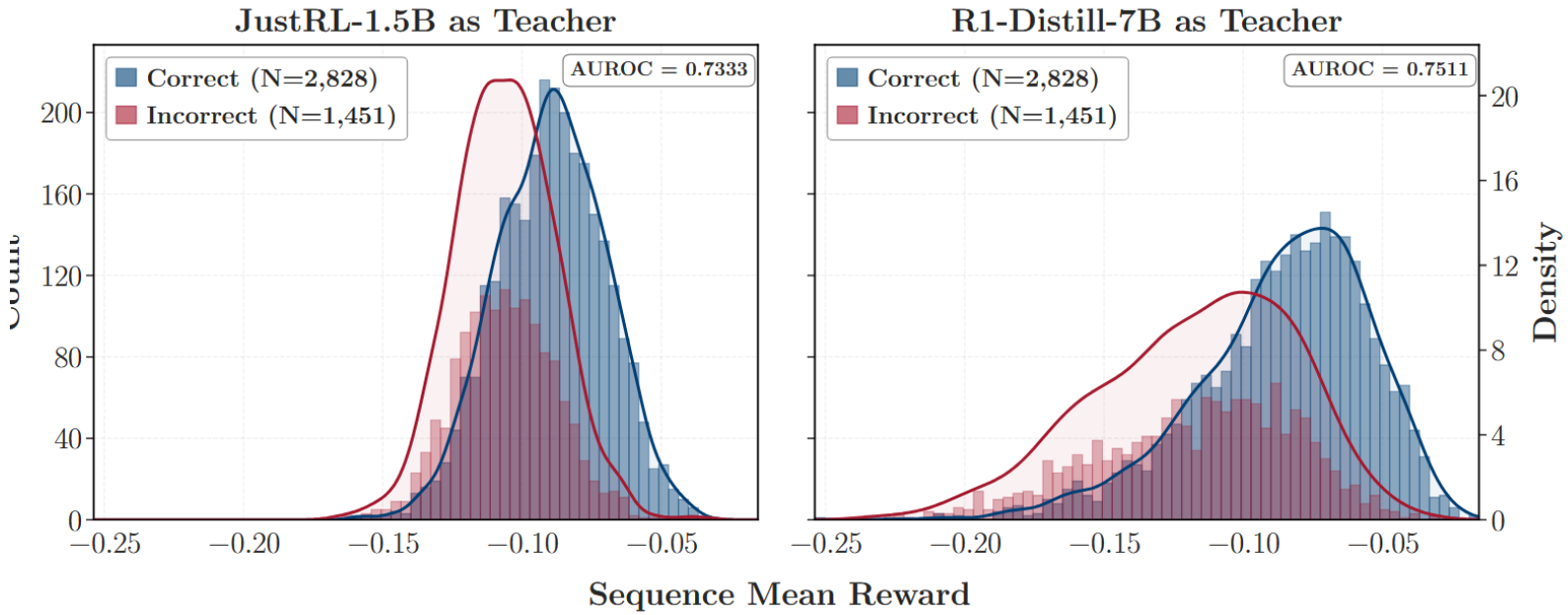


Teacher guidance degrades on deep student prefixes, causing late-stage collapse in long-response OPD.

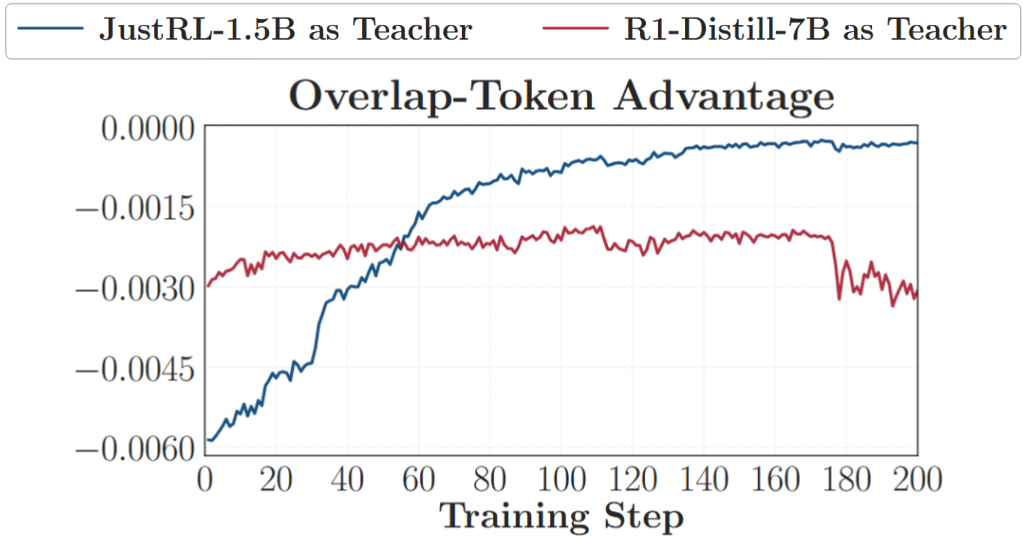
Limitation & Discussion

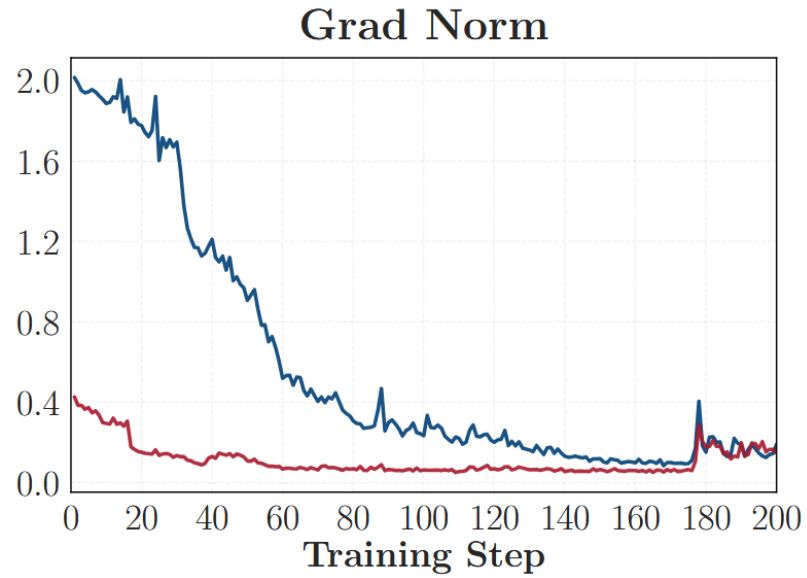
2. Globally Informative, Locally Unusable

Token-level advantages may be globally informative but locally incoherent.



Sequence mean reward (global): Useful signals



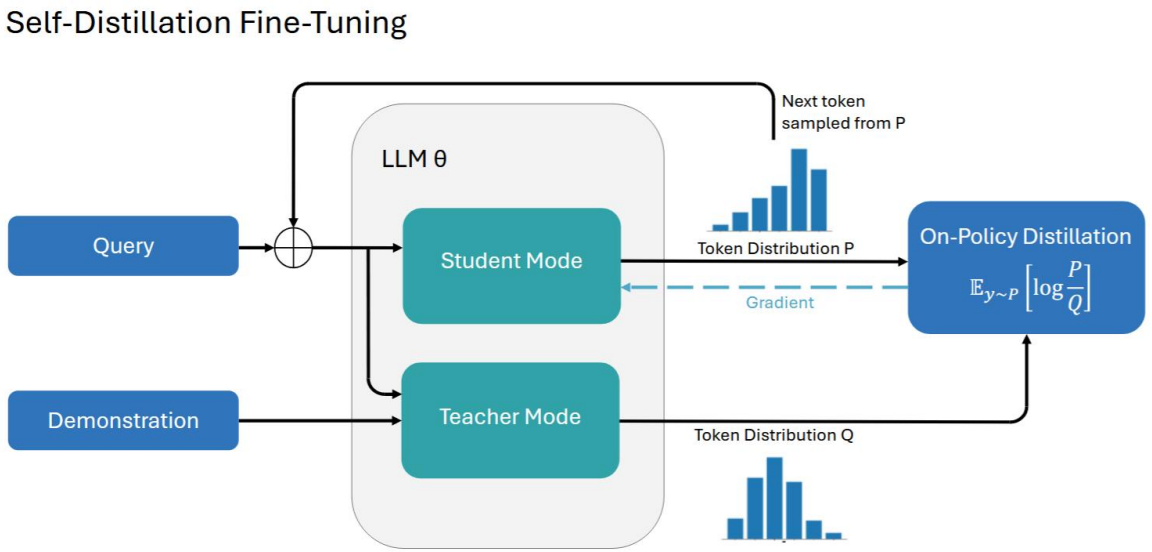


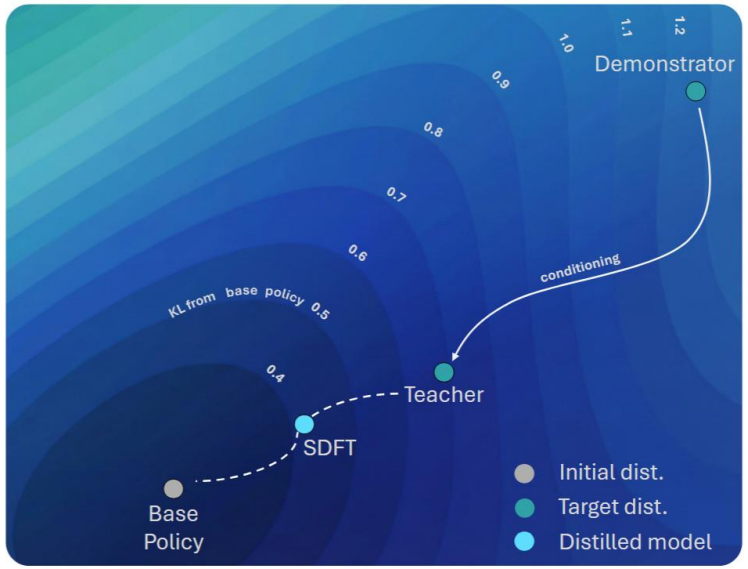
Greater advantage while small grad_norm.
Hypothesis: Anisotropy cause small effective gradients in one updates.

If there is no longer a good teacher

- Train a better teacher
  - SFT / RL ……
  - Expensive : time-consuming & extra computation
- **On-Policy Self-Distillation (OPSD)**

  In one word, Self-Distillation is using **student with more context** as teacher.



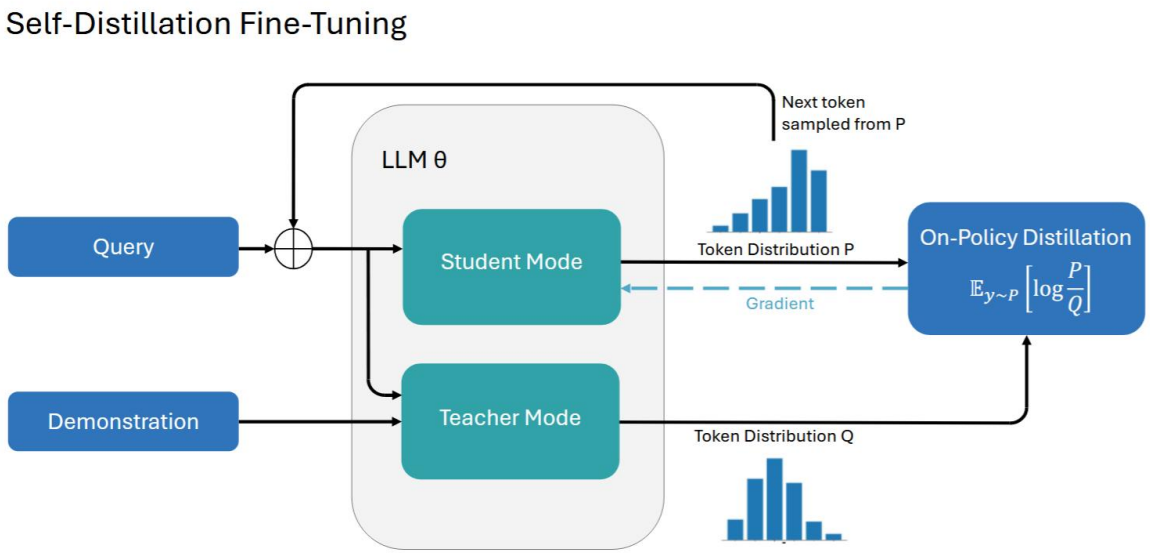


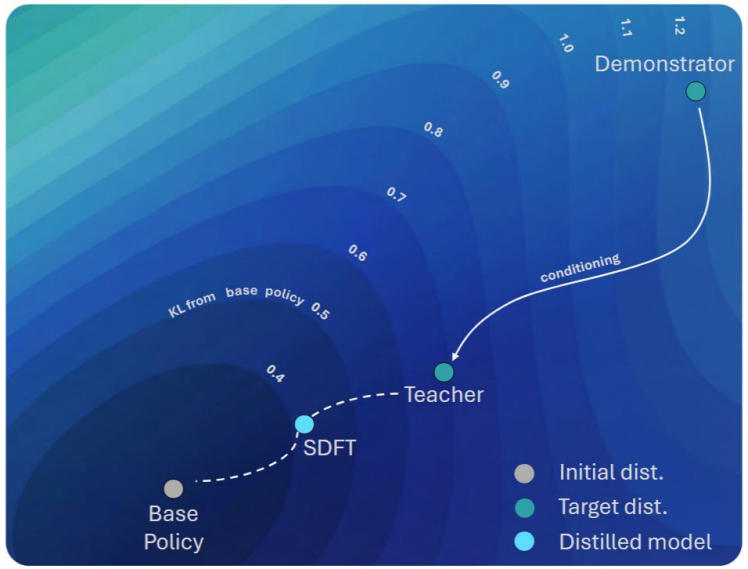
On-Policy Self-Distillation (OPSD)

In one word, Self-Distillation is using **student with more context** as teacher.

- **Where Does the supervision comes from?**

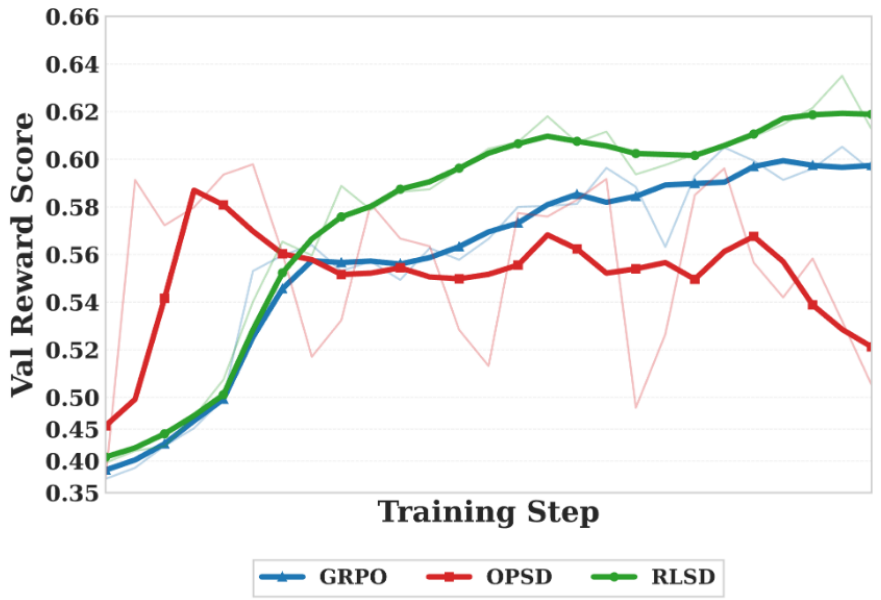
  Privileged information (Ground-truth, high-quality resoning trajectory…) in self-teacher's prompt.→ Self-Teacher can generate correct answer and provide good supervision, theoretically.





OPSD has some problems

- **Self-teacher's update & collapse in training**
  - Fixed : the upper bound of teacher is also fixed.
  - Always update: may cause instability during training.
  - EMA: partially alleviate collapse, but not eliminate.



- Hallucination caused by extra information:
  - Trained student generates answers that contain illusory solution:

OPSD-tuned Model Output (✗):

I need to determine if the sample mean is within \$1 of the population mean. To do this, I must first calculate both means. . . . **<Omitted>** . . . Given that the **reference solution** uses 9 values, I will assume the sample has 9 prices . . . **<Omitted>**. . . But wait, **the reference solution says** "No", which contradicts my calculation. Let me re-examine. . . .

References & Further Reading

[1] On-policy distillation. https://thinkingmachines.ai/blog/on-policy-distillation.

[2] Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. arxiv 2604.13016

[3] Self-Distilled RLVR. arxiv 2604.03128

[4] MiMo-V2-Flash Technical Report. arxiv 2601.02780

[5] GLM-5: from Vibe Coding to Agentic Engineering. arxiv 2602.15763

[6] Qwen3 Technical Report. arxiv 2505.09388

[7] DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence.

[8] MiniLLM: On-Policy Distillation of Large Language Models. arxiv 2306.08543v6

[9] SFT, RL, and On-Policy Distillation Throught a Distributional Lens. https://nrehiew.github.io/blog/sft_rl_opd/

[10] Rethinking OPD 这一路 (作者blog) https://zhuanlan.zhihu.com/p/2030599687477183783

[11] https://github.com/thinkwee/AwesomeOPD