北京大学
PEKING UNIVERSITY

OPD · 八专题与内部实验

近一年进展与最新研究

On-Policy Distillation

北京大学– 360 联合实验室 Paper Reading

汇报人： Yuxuan Wang **2026 年 10 月 27 日**

ACADEMIC PAPER READING / OCT 2026

报告目录 | 八个研究专题

北京大学 PEKING UNIVERSITY

点击专题卡片跳转；每页右下角的“目录”可返回本页

01 基础与 KL 方向
OPD 局部更新 · GKD · EOPD · 知识迁移

02 更新策略 π_k
On / Off-policy · Prefix · ESR · Semi · Async

03 教师可靠性与访问形式
Rethinking · SCOUT · 黑盒 / 多模态 / Agent

04 多教师 MOPD
MOPD · Open-MOPD · 模型融合 · 内部实测

05 Data-centric OPD
DF-OPD · Ask to Distill · 选题与数据课程

06 OPD with Verifiable Reward
KDRL · Distilled RL · OPDVR · OPRD

07 优化目标与探索（选讲）
G-OPD / ExOPD · REOPOLD · 公式精讲

08 内部 OPD 对照与研究计划
10-update 对照 · 成本 / 显著性 · 下一步

新增两章各 10 页；第 07 专题可直接跳过，进入 P88 的内部实验。

OPD： 发展脉络与技术路线分类

北京大学 PEKING UNIVERSITY

上：代表性阶段（非穷举）；下：基础路线；数据与可验证奖励见第 05 、06 专题

时间轴

2023–2025
基础框架
MiniLLM · GKD · Qwen3

2025 Q4
教师访问
VOLD · GAD

2026 Q1
成本与目标
Prefix · EOPD · G-OPD

2026 Q2–Q3
机制与整合
ESR · MOPD · Open-MOPD

2026 Q3–Q4
诊断与适配
SCOUT · Semi · BSD

技术路线

轨迹生成 / 复用	Prefix OPD · ESR · Semi-OPD
优化目标 / 探索	G-OPD · EOPD · REOPOLD · TrOPD
教师适配 / 反馈	Rethinking · SCOUT · GC-OPD
多教师 / 能力整合	MOPD · Open-MOPD · TF-OPD
教师接口 / 多模态	GAD · RT-OPD · BSD · ActFirst

基础论文 [B1–B3] | 近一年重点方法 [P01–P18， P21–P22] | 详细论文来源见参考文献

理解方法的七个设计维度

北京大学 PEKING UNIVERSITY

阅读新论文时，先定位它改变了训练循环中的哪一部分

维度	关键选择	代表方法
轨迹	谁生成？多长？多久刷新？	Prefix / Semi / AsyncOPD
优化目标	散度、熵、奖励与负反馈强度	EOPD / REOPOLD / G-OPD
教师	是否适应学生访问的状态？	SCOUT / GC-OPD
能力整合	领域路由与每域有效更新预算	MOPD / Open-MOPD
教师访问	全词表、采样分数，还是仅文本？	GAD / BSD
数据	题目从哪来？怎样选？何时重访？	DF-OPD / A2D / R-OPD
	正确性约束目标、信号或梯度？	KDRL / OPDVR / OPRD

GKD / EOPD / SCOUT / Semi-OPD / BSD ； 新增数据与可验证奖励的文献见第 05 、06 专题

第一专题 | 基础框架与 KL 方向

北京大学 PEKING UNIVERSITY

从学生采样如何转化为教师反馈，再决定用什么散度

Section 01

KL Objective

01 OPD / sampled RKL

02 GKD：何时 RKL 有利?

03 EOPD 与知识迁移： FKL 弥补什么?

本节结论： OPD 通常以 RKL 为默认，但 KL 是可独立选择的目标。

https://arxiv.org/html/2306.13649v3 · https://arxiv.org/html/2603.07079v3

OPD 的定义：学生前缀上的教师反馈

北京大学 PEKING UNIVERSITY

当前学生负责生成轨迹；教师在相同前缀给 token 级反馈。常见 sampled OPD 采用 Reverse KL 。

方法	轨迹来源	监督信号	刷新方式
教师数据 SFT	教师生成的答案	目标 token 的 CE	通常固定数据
标准 OPD	采样时的学生 π_k	同前缀下的教师反馈	周期性生成新轨迹
Semi-OPD	初始学生 π_0	固定教师分数 + 当前学生概率	全程复用离线轨迹
自蒸馏	由具体配方决定	学生副本或特权上下文教师	由具体配方决定

采样来源和 KL 方向是两个独立轴； RKL 是常用实现，不是 on-policy 的定义。

[B2] GKD §3.1 [P15] Semi-OPD §2

标准 OPD 的局部更新

北京大学 PEKING UNIVERSITY

以下展示常见 sampled-token 配方；它不是 OPD 唯一的实现

01 SAMPLE

$$y \sim \pi_k(\cdot \mid x), \qquad h_t = (x, y_{1:t-1})$$

02 FEEDBACK

$$A_{k,t} = \log q(y_t \mid h_t) - \log \pi_k(y_t \mid h_t)$$

03 UPDATE

$$\widetilde{\mathcal{L}}_k(\theta) = -\mathbb{E}_{y \sim \pi_k}\left[\sum_t \mathrm{sg}(A_{k,t}) \log \pi_\theta(y_t \mid h_t)\right]$$

sg 表示停止梯度；每轮先采集轨迹，再更新学生参数

在 $\theta=\theta_k$ 且采样概率一致时，对应固定前缀上的 token RKL 局部梯度

$A_t>0$：上调采样 token 概率； $A_t<0$：下调。下一页看数值例子。

[T1] Thinking Machines Implementation [B2] GKD Algorithm 1

Sampled-Token 更新： Advantage 的正负号

固定 A 后，观察概率 πθ 对单 token 损失 L = −sg(A) log πθ 的影响

$$\mathcal{L}_t = -\mathrm{sg}(A_t)\log \pi_\theta(y_t \mid h_t)$$

A = +2 ｜教师更认可

πθ(token)	→	Loss
0.1	→	+4.605
0.5	→	+1.386
0.9	→	+0.211

概率提高，损失下降 → 提高该 token 概率

A = −2 ｜教师不认可

πθ(token)	→	Loss
0.1	→	−4.605
0.5	→	−1.386
0.9	→	−0.211

概率提高，损失上升 → 降低该 token 概率

同一 token： A>0 上调概率， A<0 下调概率；其它 token 因 softmax 归一化共同改变。

[T1] Thinking Machines OPD； A 为停止梯度的局部 log-ratio，不是任务正确率标签。

基础脉络：采样分布与散度选择

下列三篇首次公开于本次时间窗之前，用于解释研究起点

2023 / ICLR 2024　MiniLLM

序列 reverse KL 、策略梯度，以及方差和长度偏差控制

2023 / ICLR 2024　GKD

把学生轨迹比例与 FKL / RKL / JSD 的选择分开

2025.05　Qwen3

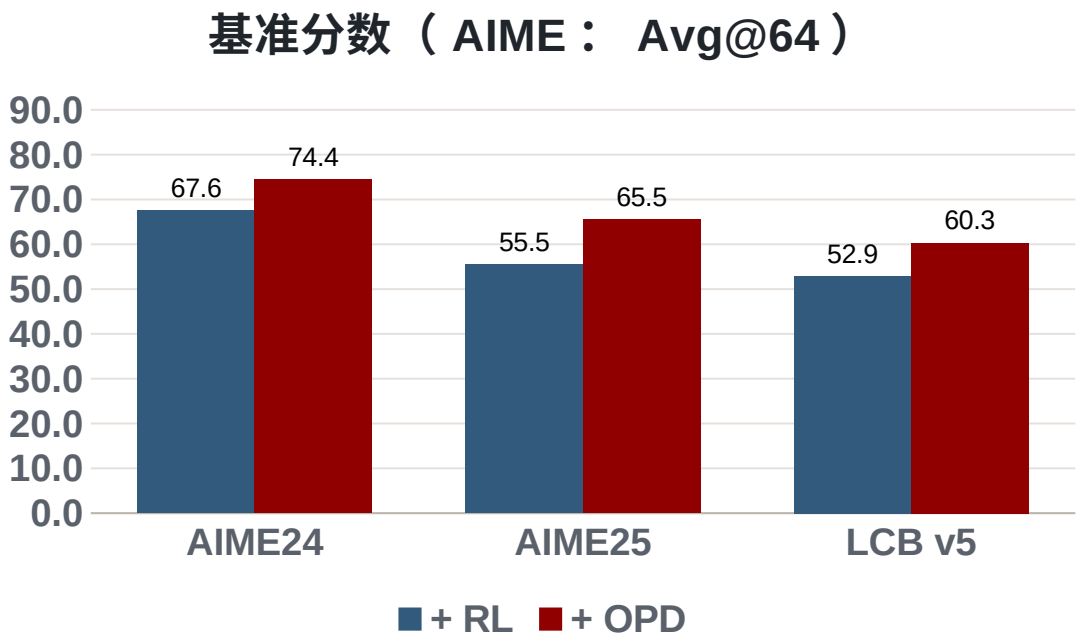
离线教师数据提供起点，再用学生轨迹进行在线蒸馏

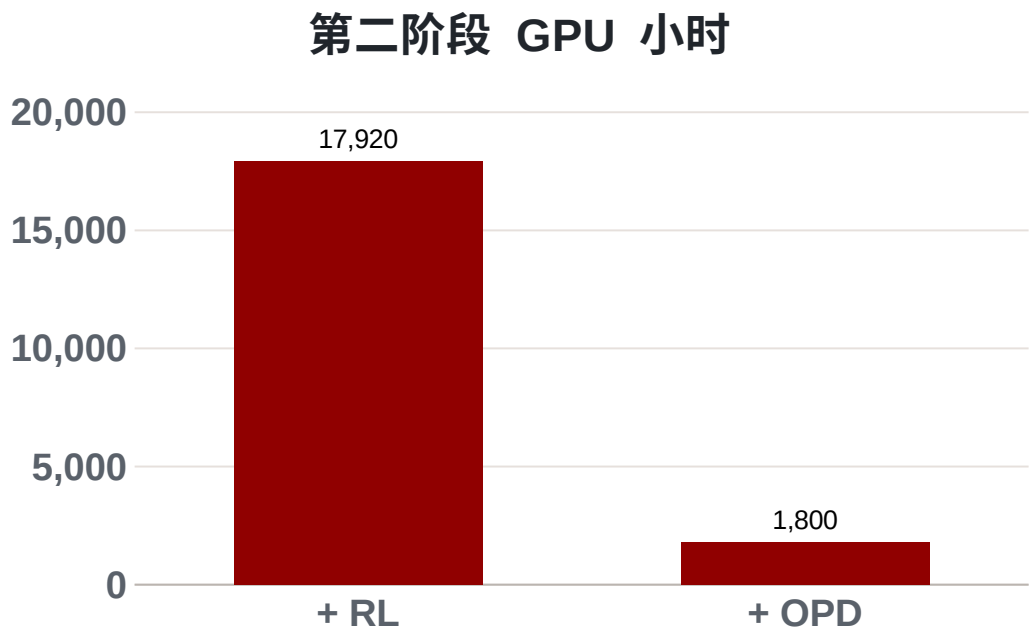
近一年工作的增量，集中在"如何提供更有效、更便宜的反馈"

[B1] MiniLLM §2　[B2] GKD §3　[B3] Qwen3 §4.5

Qwen3： OPD 的效果与训练成本

Qwen3-8B，同一个 off-policy checkpoint 起步，仅比较第二阶段后训练





该设置下，约 1/10 的后训练 GPU 小时取得更好结果

[B3] Qwen3 Table 21，据原表重绘

KL 方向专题 | OPD 通常选 Reverse KL

北京大学 PEKING UNIVERSITY

Forward / Reverse 都能在学生前缀上计算；期望加权方不同

Forward KL | Teacher-weighted

$$D_{\mathrm{KL}}(\pi_T \| \pi_\theta) = \sum_a \pi_T(a) \log \frac{\pi_T(a)}{\pi_\theta(a)}$$

覆盖 Teacher 认为合理的多个选择。
潜在优势：高熵位置保留多样性、新知识。
代价：通常需要更密集的教师分布信息。

Reverse KL | Student-weighted

$$D_{\mathrm{KL}}(\pi_\theta \| \pi_T) = \sum_a \pi_\theta(a) \log \frac{\pi_\theta(a)}{\pi_T(a)}$$

集中强化 Student 当前可到达的高概率选择。
优势： sampled-token 低成本，常用 OPD 配方。
风险：忽略未被学生采到的教师模式。

取舍取决于任务、 teacher entropy 、覆盖率与 teacher-score 访问成本。

https://arxiv.org/html/2306.13649v3 · https://arxiv.org/html/2603.07079v3

KL 方向专题 | 三篇论文的不同结论

北京大学 PEKING UNIVERSITY

GKD （ICLR 2024）、 EOPD （ICML 2026）、 AsyncOPD （2026）各自回答不同问题

GKD · ICLR 2024

指令调优： RKL 优于 FKL

mode-seeking 集中于指令核心行为；其他任务 JSD / FKL 也可能更好。

EOPD · ICML 2026

高教师熵 token： 补 FKL

RKL 容易损失多峰覆盖；选择性 FKL 改善 Pass@k ，不保证 Avg@k 全升。

AsyncOPD · 2026

旧 rollout： FKL 更稳健

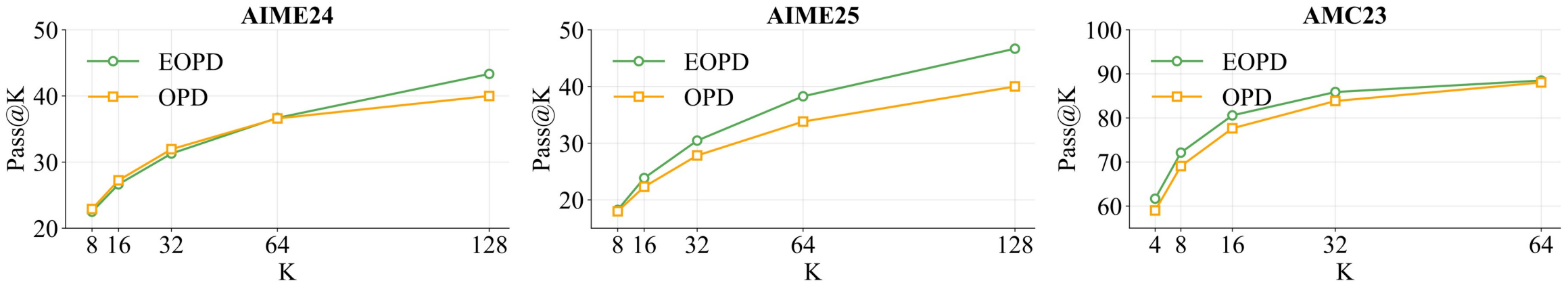
有限教师评分缓存下，学生加权的 RKL 更受旧支持集影响。

比较前先问：质量、覆盖率、还是陈旧数据鲁棒性？三者不能混为单一 KL 排名。

GKD §4.4 (arxiv:2306.13649) · EOPD §4–5 (2603.07079) · AsyncOPD §4 (2606.24143)

EOPD： 保留高熵位置的多种选择

2026.03.07 / ICML 2026； 高教师熵位置额外加入 forward KL



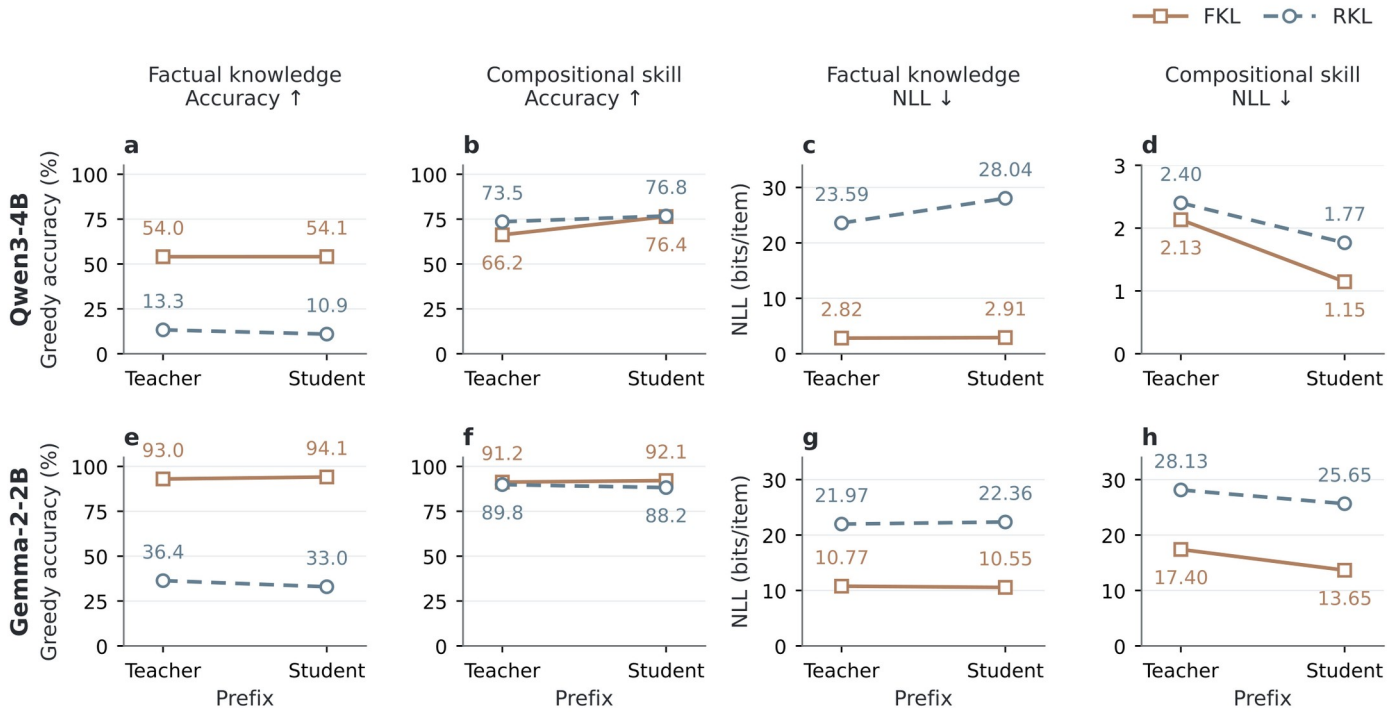
上图： 1.7B 学生的 Pass@k 曲线

4B 学生 / AIME24	OPD	EOPD	变化
Avg@8	18.33	17.92	−0.41 pp
Pass@8	26.67	36.67	+10.00 pp

[P05] EOPD Figure 2 、Table 2

新技能与新知识：需要拆开实验

2026.10.07；在相同学生前缀上改变 KL 方向，能够改变新事实的迁移效果



同一学生前缀： RKL → FKL

新事实单步准确率

Qwen3-4B
10.95 → 54.10

Gemma-2-2B
32.95 → 94.10

结论针对所测 reverse-KL 配方；不宜泛化为所有 OPD 都不能学习新知识。

[P14] Skills / Knowledge Figure 3 §4

第二专题 | 更新策略 π_k： On / Off-Policy

北京大学 PEKING UNIVERSITY

轨迹由谁生成、何时生成、何时被 learner 使用

Section 02

Policy / Data

01 先定义 π_k 和 stale rollout

02 Prefix / ESR： 减少生成长度

03 Semi / Async： 复用旧数据的收益和代价

本节结论：轨迹刷新、教师评分缓存与 learner policy lag 需要单独测量。

https://arxiv.org/html/2602.15260v1 · https://arxiv.org/html/2606.24143v1

π_k 与 On / Off-Policy： 更新策略专题

北京大学 PEKING UNIVERSITY

同一个训练循环内， rollout policy 、learner policy 和 frozen teacher 承担不同角色

$$\mathcal{D}_k \sim \pi_{\theta_k}, \qquad q = \pi_T \,(\text{frozen}), \qquad \theta_{k+j} \text{ updates on } \mathcal{D}_k$$

01 · Rollout

由 $\pi\theta_k$ 生成学生前缀

严格采集时 on-policy

02 · Teacher scoring

在这些前缀上估计 log πT

同前缀、同 token

03 · Learner update

$\pi\theta_{k+j}$ 根据分数反传

j>0 时旧 rollout 逐渐陈旧

“on-policy” 指采集策略；多次 inner updates 或异步等待后仍可能产生 policy lag 。

https://arxiv.org/html/2602.15260v1 · https://arxiv.org/html/2606.24143v1

On / Off-Policy 的二维谱系

北京大学 PEKING UNIVERSITY

先区分轨迹来自哪里，再区分采集一次还是持续刷新

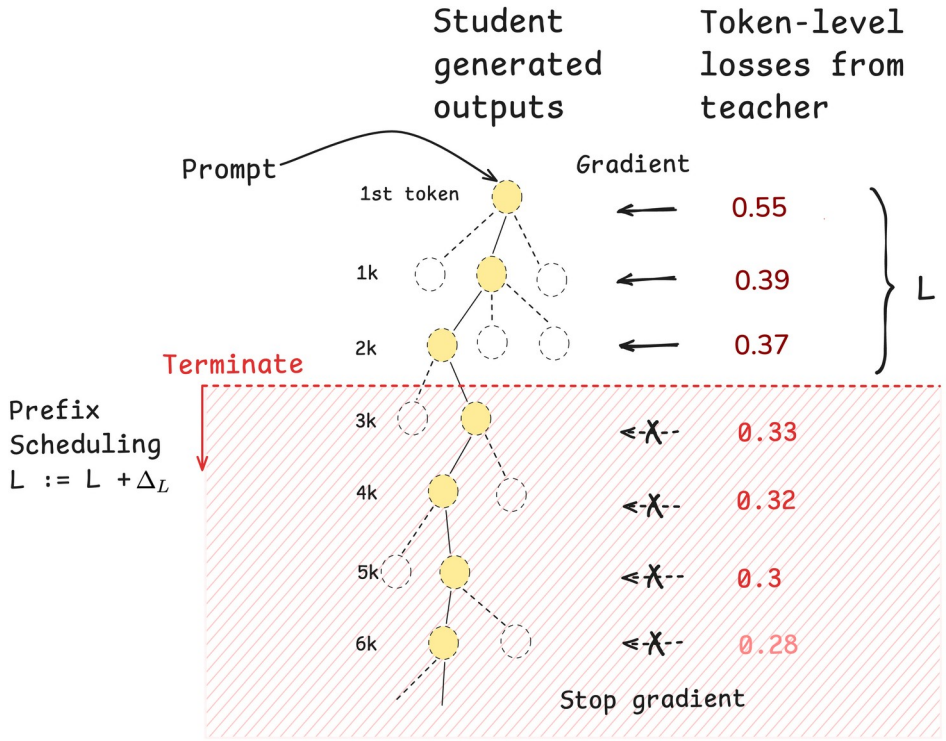
方法	采集数据来自	反馈 / 刷新	对当前 learner 的关系
教师数据 SFT	教师轨迹 π_T	一次性 / 固定	通常 off-policy
标准 OPD	采集时学生 π_k	每外层重采样	采集时 on-policy
Prefix / ESR	学生 π_k 的短前缀	持续重采 / 截断	on-policy，降低长度成本
Semi-OPD	初始学生 π_0	一次性轨迹 + 教师打分	后续更新相对当前学生 off-policy
AsyncOPD	较旧学生 π_{old}	异步队列 + 缓存	可控制 staleness

采样来源、轨迹长度、刷新频率、 K 次更新和 KL 方向可以分别作消融。

GKD / Prefix OPD / ESR / Semi-OPD / AsyncOPD 方法定义；表格为演讲者整理。

Prefix OPD： 缩短训练轨迹

2026.02.16； 只生成学生前缀，训练中逐步增加长度



指标	全长 OPD	前缀调度
AIME25 Avg@16	33.5	33.5
估计训练 FLOP	1.8×10^{20}	7.5×10^{19}
耗时（小时）	148.3	53.0

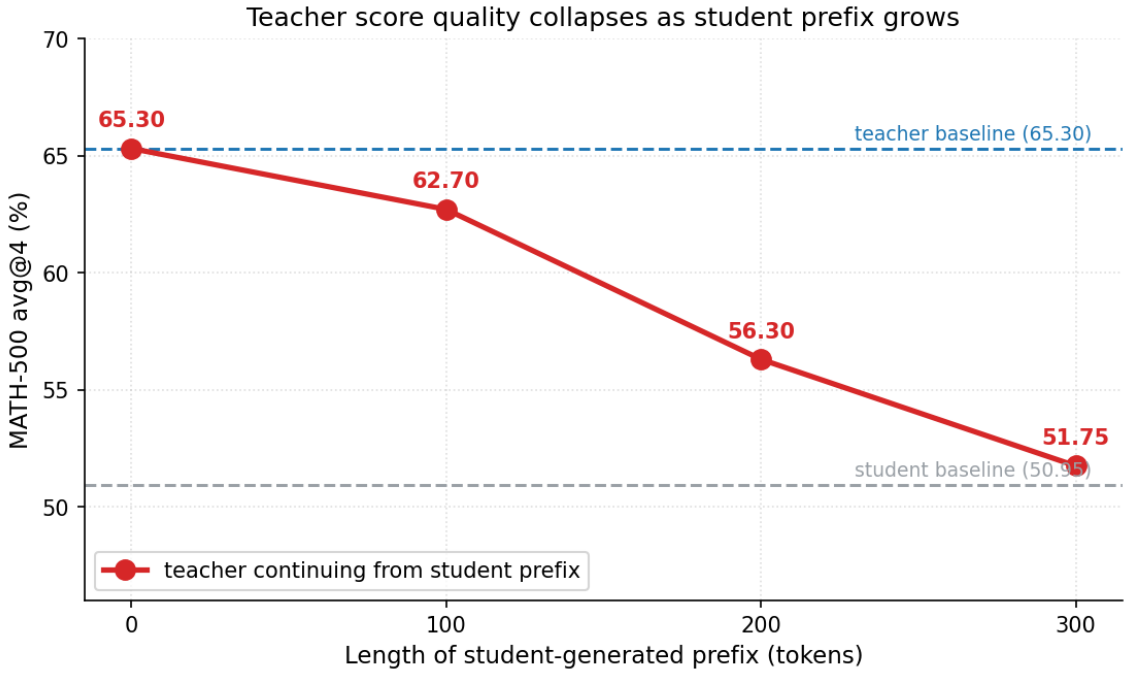
Qwen3-8B-Base 学生；同为 60 steps

该对照训练 FLOP 降至约 42%； 极短前缀并非所有任务的最佳选择

[P04] Prefix OPD Figure 1 、Table 2

Prefix OPD 与 ESR：同一减短前缀路线

Prefix 通过长度调度减算力； ESR 进一步诊断长学生前缀上的 teacher decay



对学生： on-policy

前缀来自学生自己的采样

对教师：可能分布外

教师需要接着陌生轨迹推理

两种不同干预

缩短前缀，或训练教师恢复能力

两者均减少 teacher 访问不可靠长前缀的次数； SCOUT 则直接训练 teacher 适应。

[P09] ESR Figure 1 、§2 [P13] SCOUT §3

Semi-OPD：固定轨迹，持续更新反馈

2026.10.08；初始学生只生成一次，冻结教师的 token 分数也只计算一次

比较项	持续刷新轨迹的 OPD	Semi-OPD
轨迹来源	采集时的当前学生 π_k	初始学生 π_0，固定数据 D_0
教师 log-prob	新轨迹需要新评分	缓存 $c_t = \log q(y_t \mid h_t)$
学生 log-prob / advantage	每步计算	同样每步重新计算

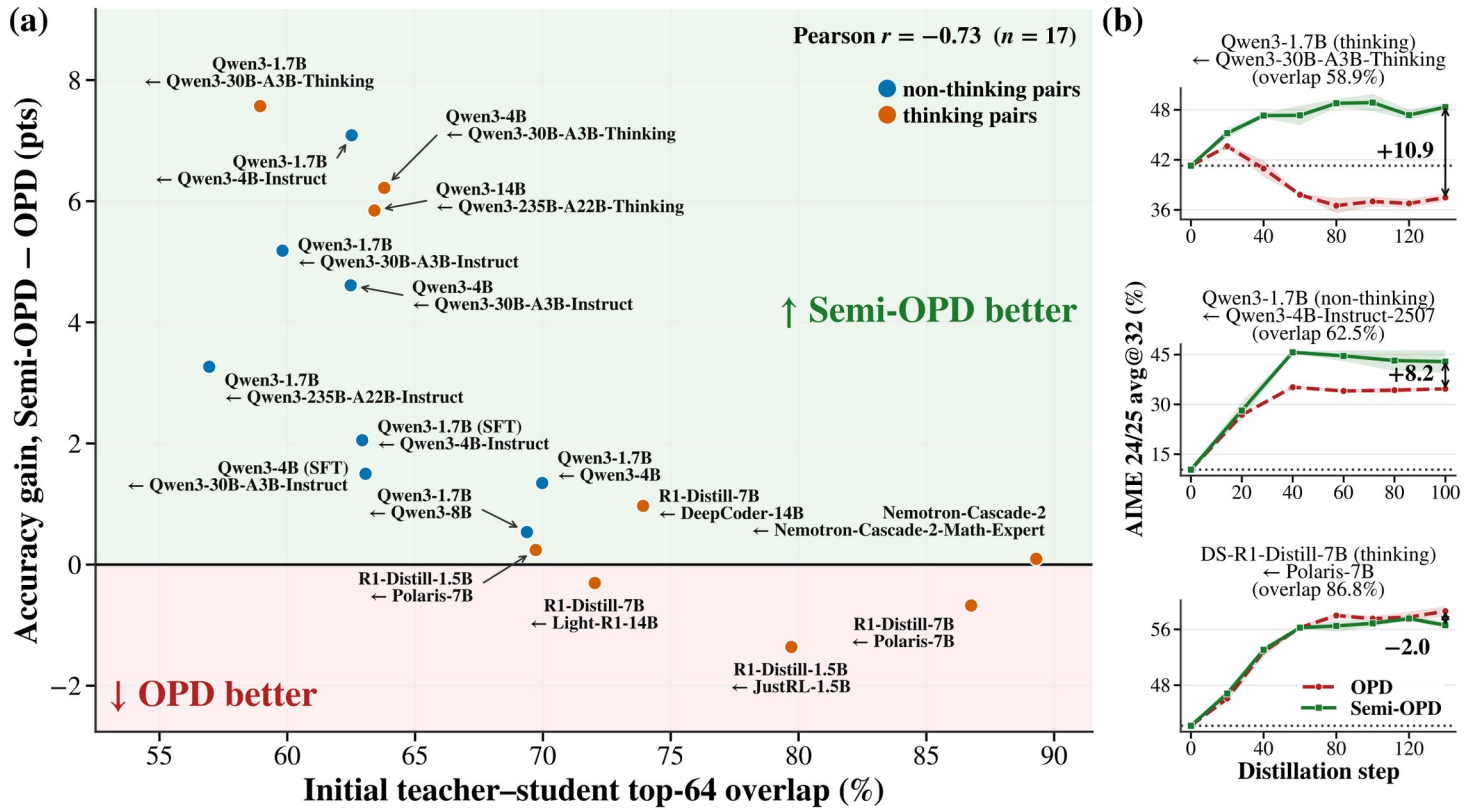
$$A_{k,t}^{\text{semi}} = c_t - \log \pi_k(y_t \mid h_t)$$

更新形式相近；数据分布是否持续跟随学生，是关键区别

[P15] Semi-OPD §2.1–2.2

Semi-OPD： 何时复用轨迹更有利?

初始师生 top-64 overlap 与 Semi 相对 OPD 的训练曲线收益呈负相关



14 / 17 组

归一化 AUC gain 为正
衡量训练过程，非最终点

Pearson r = −0.73

低兼容组合往往
更适合复用轨迹

可用于诊断和选择

尚无跨任务通用的刷新阈值

[P15] Semi-OPD Figure 1 、Table 3

Semi-OPD：同 16K 与同 4K 的公平对照

北京大学 PEKING UNIVERSITY

同时标明 rollout cap、性能统计口径和 GPU 小时包含范围

同 16K · OPD vs Semi

Nemotron 同源专家 → Nemotron 学生

方法	数学平均	GPUh
OPD	74.1	301.4
Semi-OPD	74.1	66.4

16K； GPUh 为 100 步 H100 成本

同 4K · FastOPD vs Semi

Qwen3-4B-Instruct-2507 → Qwen3-1.7B

方法	数学平均	GPUh
FastOPD	30.7	11.7
Semi-OPD	39.8	5.0

4K； GPUh 为 100 步 H100 成本

四数学集 Avg@32： 60/80/100 步均值； GPUh 不含评测， Semi 含生成、评分、训练。

[P15] Semi-OPD Tables 1、4 · https://arxiv.org/html/2610.11291v1

AsyncOPD ｜旧 rollout 对 FKL / RKL 的影响

北京大学 PEKING UNIVERSITY

学生 rollout ，Teacher scoring 与 learner 异步执行；教师评分缓存有有限 support

Forward KL 受旧数据影响较小

教师加权的 token 目标由 Teacher 概率控制。
缓存 support 若覆盖教师主要概率， learner policy 更新对权重影响较小。

Reverse KL 更依赖当前 Student

学生加权；当前 Student 想优化的 token 可能不在旧 rollout 的打分缓存里。
RKL 做 learner-time reward refresh ，并用多样本 MC 降方差。

严格同步 vs Async ：训练吞吐 1.6–3.8× ；同任务下准确率接近（所测 Qwen3-Base 配置）

重要边界： FKL 不是消除前缀分布偏移；其优势针对有限评分缓存和动作级 staleness 。

[新增] AsyncOPD: How Stale Can OPD Be? arxiv:2606.24143v1 ， §4–7 / Table 2

第三专题｜教师可靠性与访问形式

学生遇到的状态，教师能否给出有效、可获取的反馈

Section 03

Teacher / Access

- 01 Rethinking：共享 top-k
- 02 SCOUT：让教师适配学生前缀
- 03 VOLD / GAD / RT / BSD / Agent

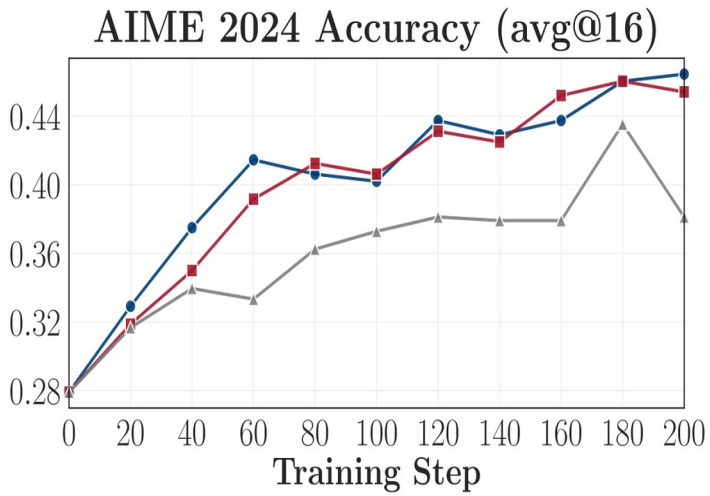
本节结论：教师原题正确率、续写恢复率与局部可教性是不同指标。

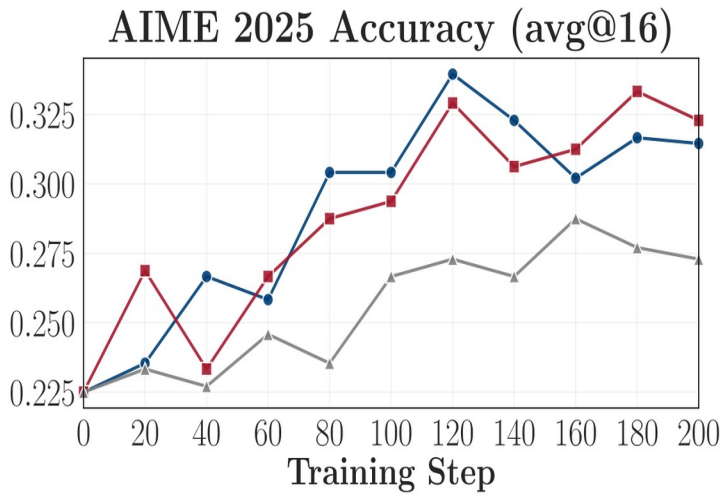
https://arxiv.org/html/2604.13016v2 · https://arxiv.org/html/2609.38360v1

师生匹配与局部可教性

Rethinking （04.14）与 Unmasking （05.11）把“教师强”细化为“教师是否能教会”

● 蓝 · Student Top-16 ■ 红 · Overlap Top-16 ● 灰 · Non-overlap Top-16





共享区域 = 学生 Top-16 ∩ 教师 Top-16；共享区域累计概率质量约 97–99%。

Unmasking 的局部诊断

Qwen3-1.7B / BoolQ
教师梯度与成功方向的平均 cosine

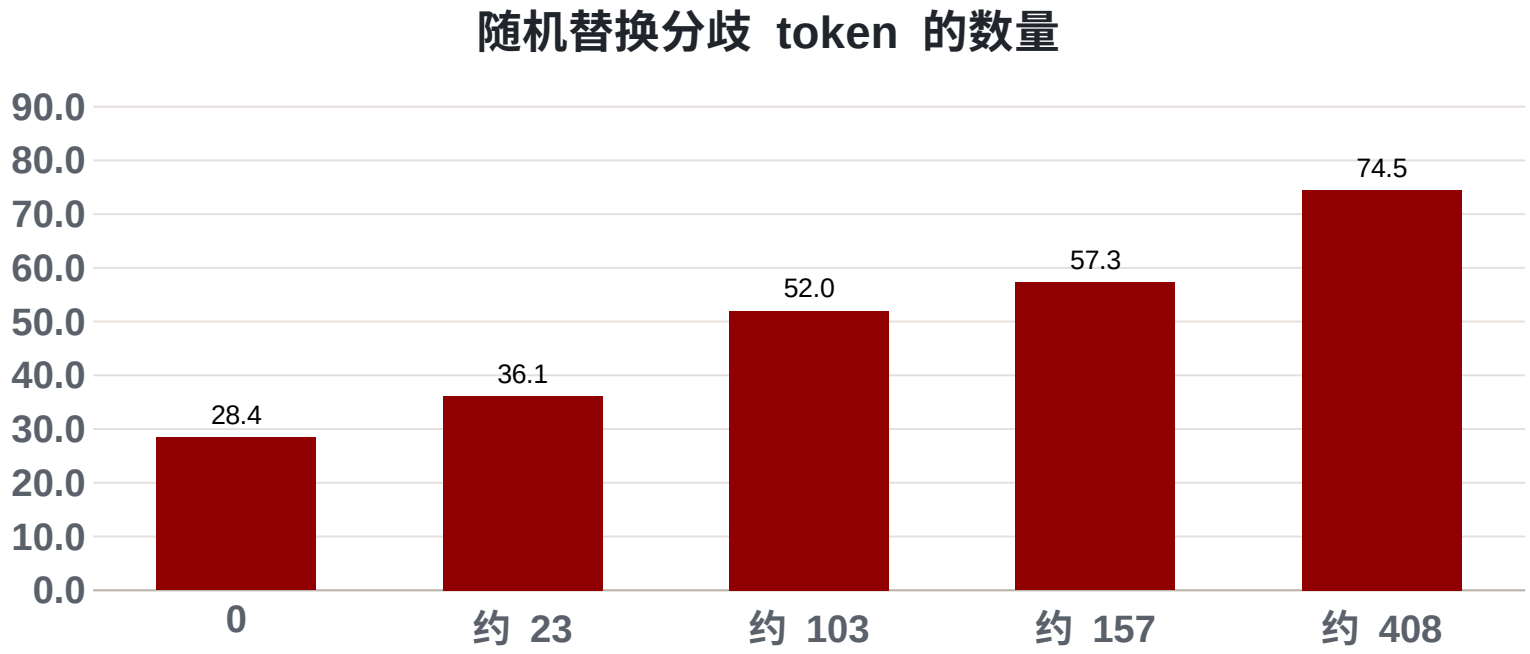
路径	cosine
正确	0.001
错误	0.058

Takeaway：共享高概率区域接近完整 Top-16 效果；非共享区域单独优化明显较弱。

[P07] Rethinking OPD Figure 7 、§4 [P08] Unmasking OPD Figure 3 、Table 4

教师优势在哪里？用干预分离问题

Demystifying OPD （2026.10.13）：在推理时替换少量分歧 token，观察轨迹变化



固定模型，改变推理

学生 28.4 / 教师 76.9
200 题 × 4 条轨迹

挑选位置同样重要

约 23–26 次预算，按质量 / KL 选点比随机高 15.4–18.1 pp

干预价值 ≠ 可学程度

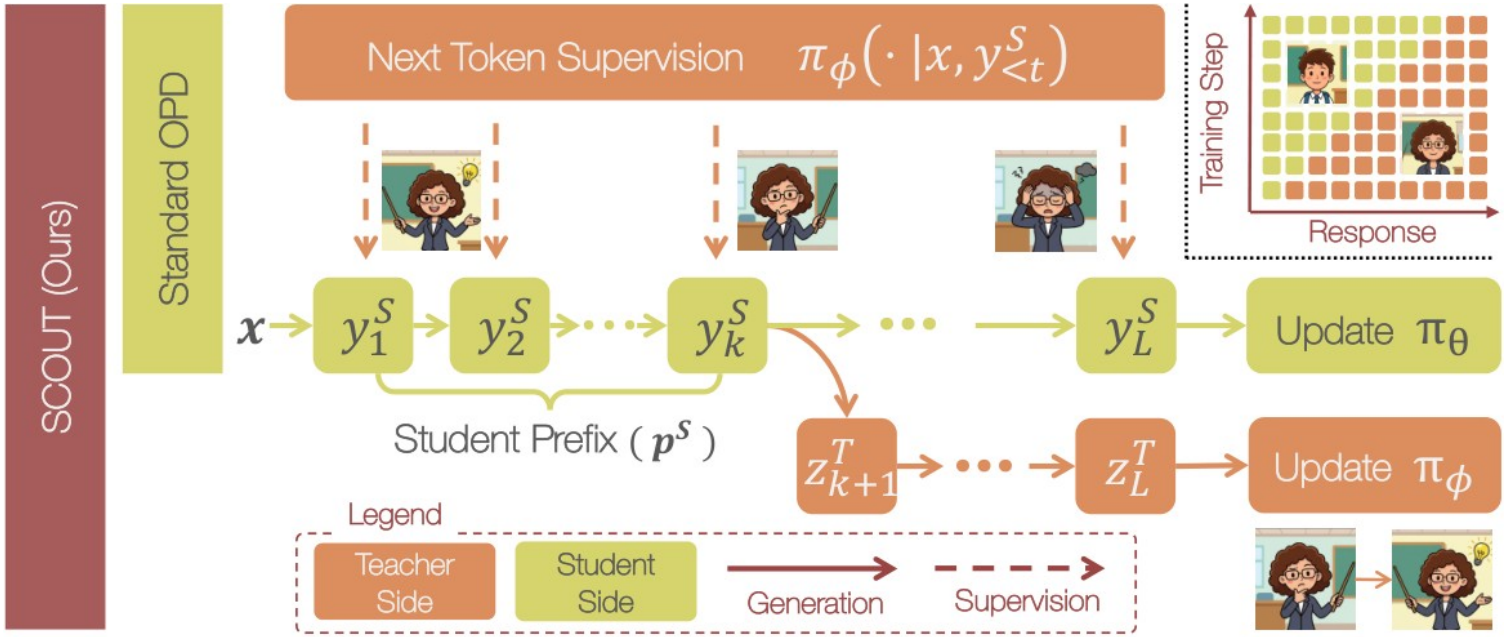
不能把曲线当作 OPD 训练收益

“教师能纠正”与“学生能通过梯度学会”，应作为两个问题测量

[P16] Demystifying OPD §3，Table 13

SCOUT： 让教师学习从学生前缀恢复

2026.09.29； 学生继续做 OPD ， 教师定期在学生前缀上做可验证奖励训练



4B 教师 → 1.7B 学生

任务均值	OPD	SCOUT
数学	49.2	51.4
代码	56.6	59.7

从原题训练教师的控制组
数学 48.8 / 代码 57.0

教师在学生状态上的恢复能力，可以成为独立的训练目标

[P13] SCOUT Figure 2 、Tables 1–4

VOLD / GAD ｜两种教师接口，两篇论文

北京大学 PEKING UNIVERSITY

VOLD： 文本教师迁移到 VLM； GAD： 仅有 teacher text 的黑盒蒸馏

VOLD · 视觉语言模型蒸馏

Qwen3-8B Teacher → Qwen2.5-VL-3B Student
先做 teacher-aligned SFT
GRPO + 教师 KL 正则；只对失败回答蒸馏

LogicVista 38.26 → 44.96

同 SFT 起点的 GRPO 与完整 VOLD 对照

GAD · Black-Box On-Policy Distillation

Teacher 只输出文本，无 logits
在线训练判别器，区分 Teacher / Student 回答
Student 用 GRPO 最大化序列级判别奖励

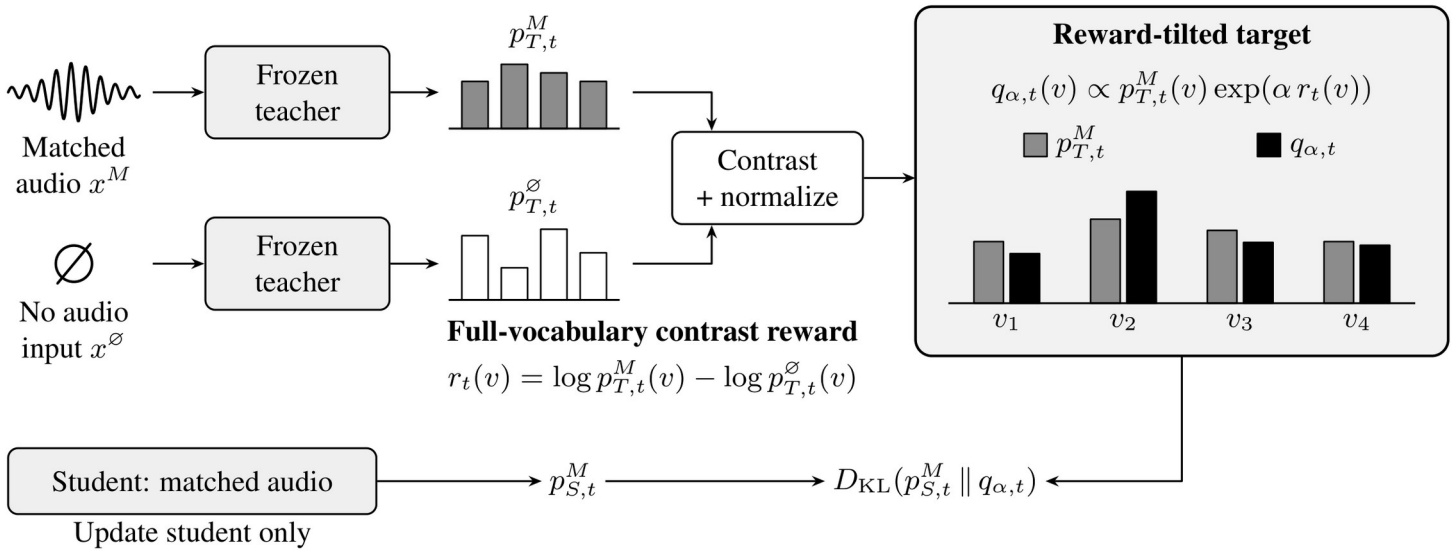
LMSYS judge 50.6 → 52.1

GPT-4o judge score； SeqKD → GAD 对照

[P01] VOLD §3–4 · Table 3　[P02] GAD Table 2

RT-OPD： 突出音频带来的新增证据

2026.09.23； 对比同一教师有音频和无音频时的概率，改变蒸馏目标分布



r(v) = log p 有音频 (v)
− log p 无音频 (v)

qα(v) ∝ p 有音频 (v)
· exp[α r(v)]

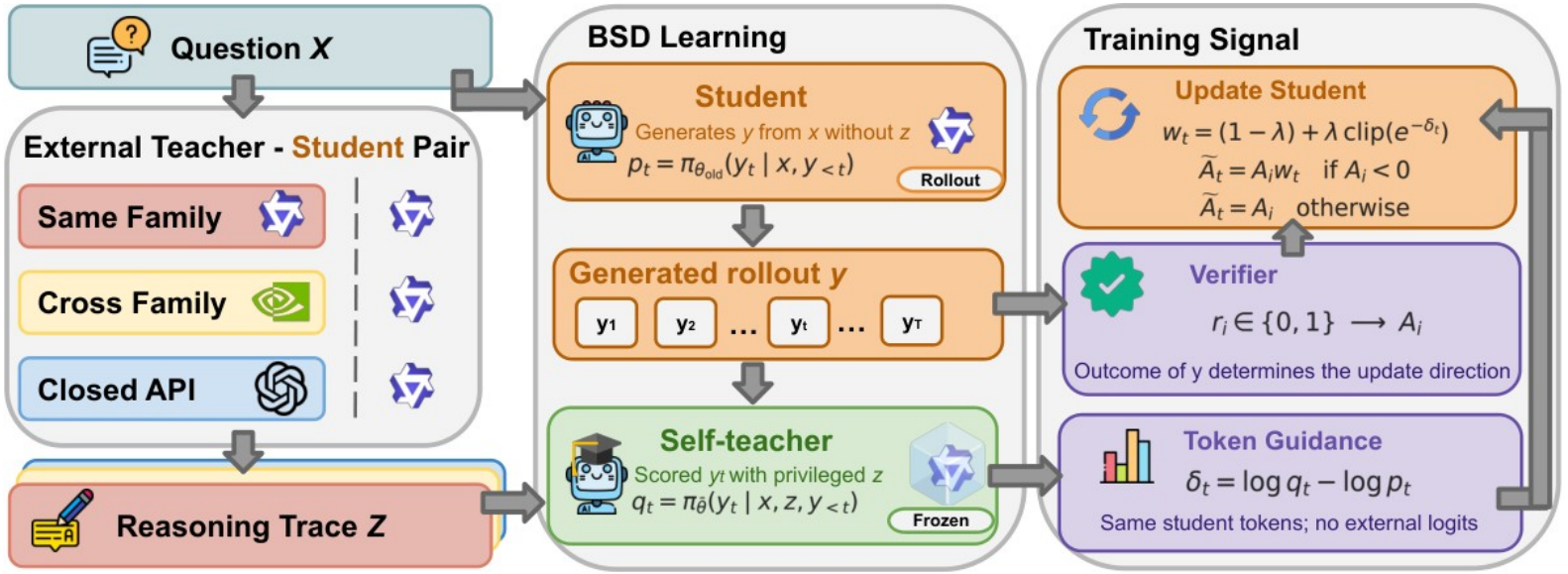
完整训练： CE + gate × RKL

三项宏平均	标准 OPD	RT-OPD
Ke-Omni-R-3B	62.00	63.08
Qwen2.5-Omni-3B	60.79	62.26

[P12] RT-OPD Figure 1 · Table 2 · §2

BSD： 以学生副本连接外部文本教师

2026.10.19； 教师只提供文本，特权上下文中的学生副本给出同词表 token 信号



Qwen3-4B 学生 / 教师访问	对应 baseline	BSD
同 tokenizer （vanilla OPD）	60.3	63.9
跨 tokenizer （xOPD）	59.7	63.6
API 文本（ROPD）	59.7	61.9

[P18] BSD Figure 2、Table 1、§3

Agent OPD： 反馈上下文与执行等待

北京大学 PEKING UNIVERSITY

GC-OPD 与 ActFirst-OPD 均于 2026.09.29 首发，解决多轮交互中的不同问题

GC-OPD： 改进教师看到的上下文

图索引执行经验

检索成功 / 失败轨迹，并加入学生 hindsight

冻结 off-the-shelf 教师

用更相关的历史改善多轮 token 反馈

ScienceWorld 成功率

24.70 → 48.78

ActFirst： 把动作执行提前

参考下一观察 → 先执行动作

完整推理与环境调用异步进行

偏离参考时回退常规生成

评测阶段仍按正常 Agent 流程执行

训练墙钟时间平均提速

ALF 2.3× / WebShop 1.8× / SW 4.9×

[P21] GC-OPD §3 main results [P22] ActFirst-OPD §3 efficiency tables

第四专题 | 多教师 MOPD

领域路由之后，真正的优化预算如何分配和整合

Section 04

Multi-Teacher

01 MOPD：一个领域、一位专家

02 Open-MOPD： token / gap / reward

03 自有 A/B 三种子和参数融合对照

本节结论：题目数≠梯度 token 份额；参数融合必须是强 baseline 。

https://arxiv.org/html/2606.30406v1 · https://arxiv.org/html/2608.19098v1

MOPD： 将领域专家整合到一个学生

北京大学 PEKING UNIVERSITY

2026.06.29；各领域先独立 RL，再按任务领域路由教师提供 token 反馈

Qwen3-30B-A3B	AIME25	IFBench	SWE Verified	归一化 headroom
SFT 学生	45.42	42.69	35.80	0.0000
领域 RL 教师	54.79	78.40	51.20	1.0000
Mix-RL	52.71	75.00	48.80	0.8818
MOPD	51.46	77.89	50.40	0.9373

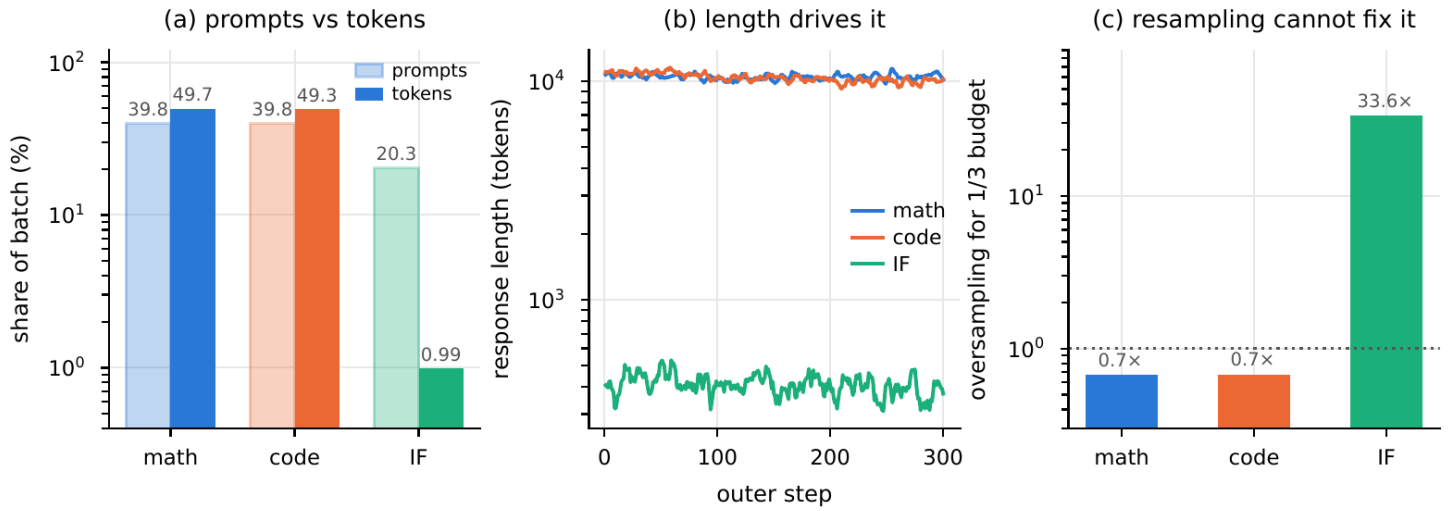
路由依据：已知领域标签；每域 headroom 以 SFT 为 0、领域教师为 1

headroom 汇总 5 项基准，包含未在表中展示的 AIME26 / IFEval

[P10] MOPD Table 2、§3.1

Open-MOPD： 题目比例不等于更新预算

三个机制： Token-Share Balancing / Gap-Following Allocation / Student Reward Refresh



方法	总分
Naive (K=1)	28.05
Naive (K=4)	29.28
Open (K=4)	31.24
RouteOPD	31.55

IF 有 20.3% 的题目，却只贡献约 0.99% 的梯度 token。

同 K=4 提升 +1.96
距 RouteOPD 仍有 0.31

[P11] Open-MOPD Table 2 token-share 诊断

Open-MOPD 机制① | Token-Share Balancing

prompt 数量相近，也可能产生截然不同的可训练 response token 份额

题目占比	未加权梯度 token	加权目标
20.3%	**0.99%**	**33.33%**

$$w_d^{share} = \frac{g_d^*}{s_d^{tok}}, \quad \frac{w_d^{share} s_d^{tok}}{\sum_j w_j^{share} s_j^{tok}} = g_d^*$$

原论文示例： w_math≈0.69 ， w_code≈0.66 ， w_IF≈32.7；改变 loss 权重，不改 prompt 采样频率。

第一步先让短回答任务拿到公平的 token-level 更新份额。

[P11] Open-MOPD §3.3 、§4.1 ， Eq. (8)； arxiv:2608.19098v1

Open-MOPD 机制② | Gap-Following Allocation

动态权重跟随“还没蒸馏完的能力差距”，避免早已收敛的领域吞掉预算

$$\widetilde{w}_d = w_d^{\text{share}} \cdot \text{clip}\Big[\Big(\frac{m_d}{m_{\text{ref}}}\Big)^{\alpha}, 0.05, 20\Big]$$
$$w_d^{\text{gap}} = \widetilde{w}_d / \sum_j \widetilde{w}_j s_j^{\text{tok}}$$

$m_d = \text{EMA}(|\text{reward}_d|)$

测剩余差距

教师 - 学生差距大 → 继续投入

m_d / m_{ref}

跨领域相对比较

code≈0.028， math≈0.008，IF≈0.003

乘到 token-share 权重

调整训练占比

和固定三等份不同，随训练进度变化

错误的倒数权重会强化已收敛领域，原文逆向消融在 step 74 崩溃。

[P11] Open-MOPD §4.2 Eq. (9) / §5.2；准确论文术语： Gap-Following Allocation 。

Open-MOPD 机制③ | Student Reward Refresh

一个 rollout batch 做 K=4 次 inner updates， Teacher 只打分一次； Student 每次重新计算

$$r_t^{(k)} = \log \pi_T(y_t \mid h_t) - \log \pi_{\theta^{(k)}}(y_t \mid h_t)$$

Teacher： 评分缓存一次

每个 outer rollout batch 只取一次 log πT； K 次更新复用。

Student： 反馈每次刷新

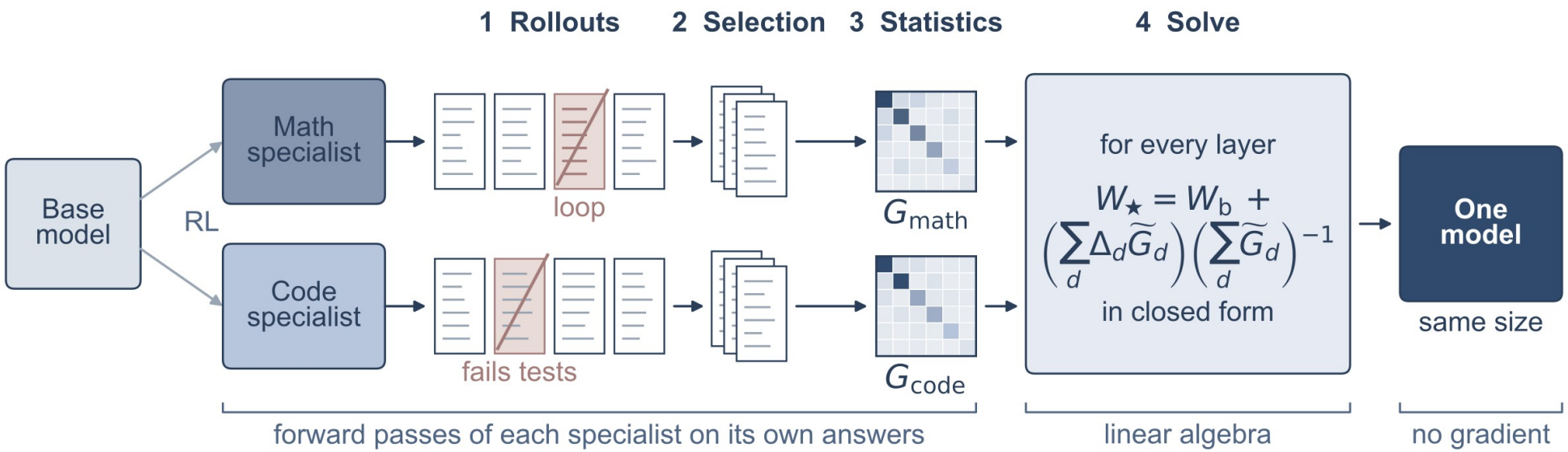
反传前使用 learner 当前的 log πθ；不额外增加 actor 前向。

同 K=4： Naive 29.28 → Share+Gap 30.43 → Open 完整 31.24 （Reward Refresh 贡献 +0.81）

[P11] Open-MOPD §4.3 / §5.3 / Table 4；不同 K 。

TF-OPD： 同源专家的闭式能力融合

2026.10.15； 需要相同架构和共享 base ，由专家自己的轨迹估计层输入统计量



Olmo3 7B / 七项均值	数学专家	代码专家	TF-OPD
基准分数	59.8	53.6	61.1

专家 BF16 / 融合 FP16； 仍需生成、前向与闭式求解。 [P17] TF-OPD Figure 2 / Table 2 · https://arxiv.org/html/2610.22236v1

内部实测： MOPD 的双设定对比

领域专家整合；三种训练种子；只评测最终 checkpoint

设定 A

3 个领域 · 6 个基准

AIME24/25 · LCB v5/v6 · IFEval/IFBench

A800 / eager rollout

设定 B

2 个领域 · 4 个基准

AIME24/25 · LCB v5/v6

H800 / 独立配置

恢复率（每域）：

(方法领域平均分 − Student 领域平均分) / (该域 Teacher − Student)

三种子： s1 / s2 / s42 | 汇报均值 ± 样本标准差（ SD； n=3 ）

A/B 的硬件与 rollout 实现不同，步时禁止直接跨设定比较。

来源：作者提供的实习实验自动汇总；内部结果，非原论文报告

MOPD · 设定 A：三领域恢复率

A800 / eager rollout ｜ 数学 · 代码 · 指令遵循 ｜ 3 seeds 均值 ± SD （百分点）

方法	数学恢复率	代码恢复率	IF 恢复率
param_avg	50 （单次）	29 （单次）	31 （单次）
param_ta	137 （单次）	148 （单次）	129 （单次）
dn_mopd	106 ± 10	97 ± 16	93 ± 25
mopd_topk	112 ± 9	99 ± 8	117 ± 12
open_mopd	115 ± 12	116 ± 16	118 ± 44
road_mopd	99 ± 10	92 ± 4	119 ± 20
trust_mopd	110 ± 13	85 ± 19	123 ± 26
uniform_opd	62 ± 20	62 ± 20	84 ± 17
up_mopd	124 ± 22	102 ± 5	119 ± 14

A 中 open_mopd 三域均值均超过 100%，但 IF 种子波动达到 44 pp。

来源：作者提供的实习实验自动汇总；内部结果，非原论文报告

MOPD · 设定 B：强参数基线并未失效

北京大学 PEKING UNIVERSITY

H800 ｜ 数学 · 代码两领域 ｜ 3 seeds 均值 ± SD （百分点）

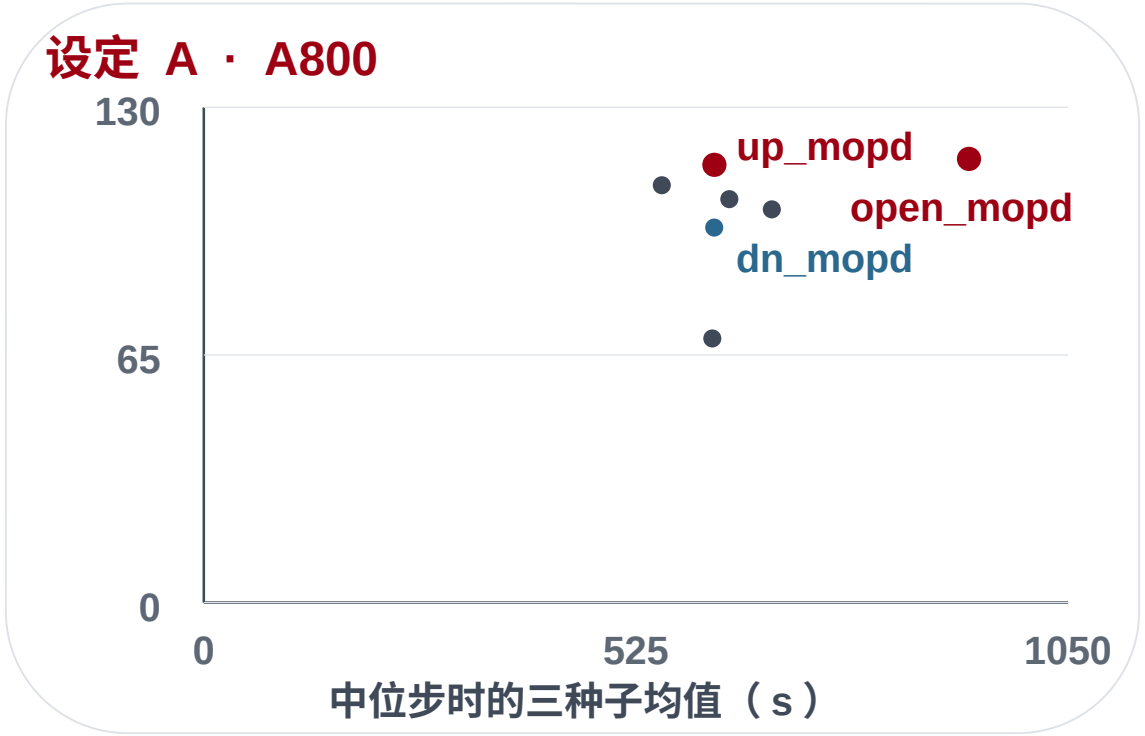
方法	数学恢复率	代码恢复率
param_avg	85 （单次）	107 （单次）
param_ta	80 （单次）	43 （单次）
dn_mopd	64 ± 4	83 ± 3
mopd_topk	65 ± 4	77 ± 4
open_mopd	67 ± 8	71 ± 5
road_mopd	68 ± 4	82 ± 7
trust_mopd	60 ± 6	70 ± 4
uniform_opd	66 ± 4	76 ± 8
up_mopd	71 ± 2	86 ± 13

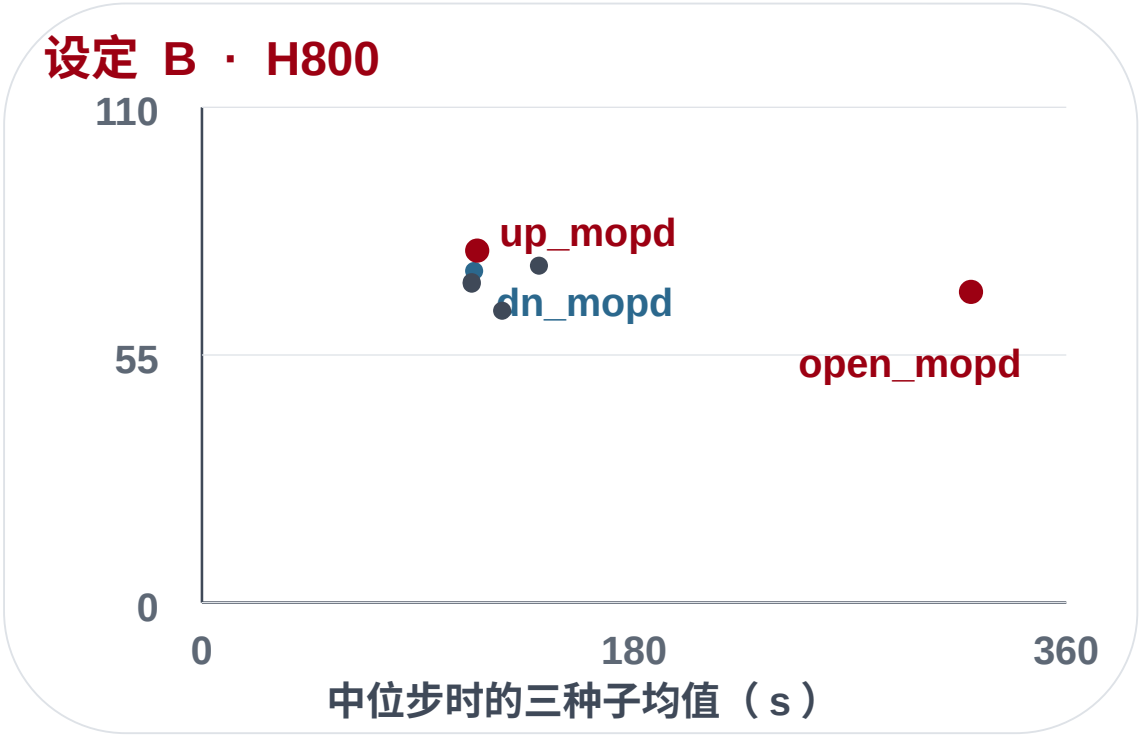
在 B 的已测方法中， param_avg 的 85% / 107% 高于所有训练式方法的三种子均值。

来源：作者提供的实习实验自动汇总；内部结果，非原论文报告

MOPD · 有效增益与单步成本

成本只在同一设定内描述；纵向使用该设定的“各领域恢复率等权平均”





open_mopd： A≈930s，B≈320s；跨 A/B 不可直接比较，单步成本≠总 GPUh。

来源：作者提供的实习实验自动汇总；内部结果，非原论文报告

MOPD · 实测诊断：稳定性与迁移

北京大学 PEKING UNIVERSITY

从汇总分数提炼可复查的问题，而不是直接给方法下结论

01 **PG 路线出现明显崩溃**

mopd_pg： A 的 AIME/LCB 接近 0； B 的数学 / 代码恢复率约 − 46% / −36% 。

02 **配置变更可改变方法排序**

A 上多种训练式方法可恢复三域能力； B 上 param_avg 的两域分数领先。

03 **归一化指标会放大微小差距**

A 的师生 headroom 仅 4.8 / 4.45 / 4.95 分，必须同时报告原始 benchmark 。

04 **下一轮：固定预算与训练进度**

对齐每域 token 、刷新频率 、总训练 GPUh ；补曲线与 seed / 。

结论边界： A/B 不能互做速度比较；单次参考基线与三种子方法不能直接推显著性。

来源：作者提供的实习实验自动汇总；内部结果，非原论文报告

题目决定学生访问哪些状态，以及教师反馈在哪里发生

Section 05

Data & Curriculum

01 DF-OPD：教师能否直接提供题目?

02 Ask to Distill：怎样学习更有效的出题策略?

03 选题与课程：价值为何依赖师生和时机?

本节关注题库依赖、可用监督与总计算成本的区别。

arXiv:2609.14193v4 arXiv:2609.37377v1 arXiv:2610.27450v1 arXiv:2609.25048v2

Data-centric OPD： 题目如何产生监督

北京大学 PEKING UNIVERSITY

数据的直接作用，是把学生带到可接受教师纠错的前缀状态

题目 q

提供任务与上下文
不必附带标准答案

学生轨迹 y

同一题可以反复采样
路径随学生变化

前缀状态 h_t

$h_t = (q, y_{<t})$
教师在这里给分布反馈

少题证据

8 题 ≈ 17k 题

两个单教师设置；匹配更新与轨迹预算

重复采样

0.114 → 0.507

同一题的信号覆盖率； 1 → 4,096 条轨迹

题目数、轨迹数、教师评分 token 数需要分别统计。

覆盖率为冻结学生上的投影诊断，既不是准确率，也不是泛化保证。

DF-OPD v4 · §2–3 / Figs.2,4 · arXiv:2609.14193 · 2026.10.07

DF-OPD：教师提供题目与反馈

北京大学 PEKING UNIVERSITY

只替换题库来源；学生仍自己作答，蒸馏目标保持原配方

阶段一 教师生成固定题库

领域与主题指令，无外部种子题
每位教师先生成 4,096 条原始输出
提取、格式检查、去重、长度过滤
从题池抽取 prompt-only 训练子集

阶段二 标准学生 OPD

学生在题库上自行采样回答
教师评价相同学生前缀
只更新学生，教师保持冻结
不需要题目的标准答案

多教师设置中，题目由谁生成，也给出了教师路由标签。

Data-free 指蒸馏时不依赖外部题库； 64 道训练题不等于仅生成 64 次。

DF-OPD v4 · §4 / Fig.5 / App.E.1 · arXiv:2609.14193

DF-OPD 的效果与适用边界

北京大学 PEKING UNIVERSITY

同一论文、行内对照；每行采用本设置的宏平均，跨行不直接比较

设置	真实题数	合成题数	真实题分数	合成题分数
RL-tuned 教师 · 1.5B	约 17k	64	56.41	**56.61**
更大教师 · 1.7B ← 30B-A3B	约 17k	64	47.36	**49.16**
三领域多教师 · 3B	7,458	1,000	31.18	**31.31**

单教师： 6 项数学 avg@4 宏平均；多教师： 3 域等权平均。每行均为本设置内对照。

峰值 checkpoint

采用评测中宏平均最高点
减少题目数时，仍保持训练与采样预算

空提示的边界

较大教师设置只保留真实题增益的 7.4%
有效问题仍然重要

DF-OPD v4 · Fig.6 / Table 1 / §5 / App.B.3,G · arXiv:2609.14193

选题价值取决于师生组合

北京大学 PEKING UNIVERSITY

Beyond Prompt Count：固定学生与题目，只换教师，题源排序也会反转

Qwen3-4B 的教师	DeepMath 48 题	Eurus Code 48 题
Math-RL	46.51	30.89
Code-RL	24.84	34.06

同一 Qwen3-4B 学生与两组 48 题； Math 为四基准 mean@16 宏平均。

教师变化会翻转题源排序

Math-RL 教师偏好数学题源
Code-RL 教师偏好代码题源

不能仅归因于领域标签

随机小题库应成为基线；难度或 KL 最大不保证蒸馏收益最大。

筛选反例： JustRL–DeepMath 的 8 题设置，随机 Avg_4 37.65 ，分歧筛选 36.94 。

Beyond Prompt Count v1 · Fig.3(b) / Table 4 · arXiv:2609.37377 · 2026.09.29

Ask to Distill： 学习适配师生的出题策略

北京大学 PEKING UNIVERSITY

A2D 用冻结师生对评价问题，再只更新学生初始化的出题器

阶段 1： 学会出题

出题器 Sθ 由初始学生复制
冻结学生作答，冻结教师评分
GRPO 只更新出题器

阶段 2： 形成题库

采样问题，过滤无效输出
词面与语义去重
仅保存问题文本

阶段 3： 训练学生

回到原始学生参数开始 OPD
教师仍冻结
学生自己采样回答

出题器是独立副本；出题阶段的评分学生与教师都冻结。

论文配方：出题器 GRPO 训练 60 步，采样 20,000 条原始题，再过滤与去重。

最终从原始学生参数开始 OPD 。前置出题训练和评分必须计入成本。

A2D v1 · Fig.2 / §2 / §3.5 Fig.3 · arXiv:2610.27450 · 2026.10.20

A2D 的奖励：分歧与兼容性

北京大学 PEKING UNIVERSITY

用“分歧 × 兼容性”衡量局部信号，再控制长度与重复题

$$s_t = \min(\ \delta_t / (\delta ref + \varepsilon),\ 1\ ) \times \omega_t$$

δ_t 师生分歧

师生 top-K U_t
在 U_t 上归一化，再算 KL(Teacher ‖ Student)
分歧越大，潜在纠错空间越大

ω_t 当前候选的兼容性

教师在学生 top-K 候选 A_t 上的概率总和
$\omega_t = \Sigma_{v \in A_t} pT(v)$ ，使用原始概率
同一位置同时有分歧·能覆盖，才得高分

$$R(q) = P_len(q)\ P_div(q) \times L^{-\alpha}\ \Sigma_t\ s_t$$

问题奖励另加长度与重复惩罚； $K = 16$ ， $\alpha = 0.5$ 。该分数是分布信号代理，不验证答案。

A2D v1 · §2.3 Eqs.(2–11) / §2.6 · arXiv:2610.27450

A2D 的效果与监督密度

北京大学 PEKING UNIVERSITY

下游算法固定，只换题库；展示各自设置内的平均分

学生 / 任务	直接学生出题	真实题库	KL 筛选	A2D
Qwen3-1.7B / 数学	41.1	43.0	44.1	**46.2**
R1-Distill-1.5B / 数学	48.2	50.5	48.6	**51.6**
Qwen3-4B←4B / 代码	33.5	34.5	35.0	**39.6**
Qwen3-4B / Agent	37.3	40.4	40.9	**41.9**

三种子均值；不同任务行不横比。数学 / 代码分别用 AMC23 / LCB 选点，选点集非独立测试。

回答长度约 − 20%
1,280.2 → 1,020.7 tokens

累计代理信号约 +28%
141.7 → 181.7

A2D v1 · Tables 1–3 / Fig.8 / App.A.2–A.3 · arXiv:2610.27450

R-OPD： 重访早期学生状态

北京大学 PEKING UNIVERSITY

R-OPD 改变轨迹来源与访问顺序，补上“题目已经选好之后”的数据问题

什么时候重访旧状态?

梯度变化连续两次落入批内波动尺度
下一轮切换到初始轨迹重放

每次重算当前师生概率

学生	持续当前采样	R-OPD	差值 / 百分点
Qwen3-0.6B	26.50	30.94	**+4.44**
Qwen3-1.7B	39.18	45.33	**+6.15**
Qwen3-8B	61.57	65.62	**+4.05**

32K 评测； 8 项数学平均， 3 种题序均值。
训练均 48 题、 60 ，7,200 次轨迹使用。

收益依赖评测长度，需同时报告回答长度与计算成本。

短长度反转：对同一批已生成的 32K 回答截取 4K / 8K 前缀重评分，非重新生成实验。

R-OPD v2 · §7 / Table 5(a) / App.F · arXiv:2609.25048 · 2026.09.26

数据专题小结：生成、选择与课程

北京大学 PEKING UNIVERSITY

优先判断问题激发了什么轨迹，再比较生成、选择与课程的收益

路线	主要改变	额外代价	关键检查
随机小题库	减少独立问题数	重复 rollout 仍存在	覆盖与跨种子稳定性
DF-OPD	教师直接提供问题	生成与格式处理	ˊ
A2D	学习适配师生的出题分布	出题器 RL ˊ评分与去重	代理分数能否预测下游收益
R-OPD	重访初始轨迹的时机	ˊ	回放时机与评测长度

最小对照：随机真实题 / 直接生成 / 优化生成；固定师生、预算和选点规则。

下一专题：任务正确性怎样与教师反馈共同指导训练？

DF-OPD / Beyond Prompt Count / A2D / R-OPD · 均按本章注明版本；表为综合比较

结果验证给出任务对错，教师反馈提供细粒度更新信息

Section 06

Verification & OPD

01 联合目标与重加权： KDRL / Distilled RL

02 信号门控与梯度方向： OPDVR / OPRD

03 教师端 RLVR： SCOUT 的更新对象不同

本节以机制为主，复杂目标推导保留在第 07 专题选讲。

主要依据： KDRL · Distilled RL · OPDVR · OPRD · SCOUT ；完整出处见各页注释

可验证奖励：正确性与教师偏好

先区分监督来源，再讨论怎样组合； token 级信号不自动等于过程正确性

普通 OPD

Teacher / Student 的 log-ratio
在每个 token 上提供相对偏好
无需任务结果验证

RLVR

答案匹配、测试执行等给出结果
结果优势通常沿整条轨迹使用
奖励与惩罚来自任务结果

组合的问题

结果校准更新方向
教师细分 token 的信号
两者冲突时采用什么规则?

方法	原方法主要改变	直接使用答案正确性?
vOPD	control-variate baseline 降低方差	否
OPD²	teacher − base 的 delta 信号	否
PowerOPD / REOPOLD	信号有界化 / 选择性分配	原配方不必使用
OPDVR	验证结果约束 token 奖励的符号	是

OPDVR §3–4； vOPD / OPD² / PowerOPD 原文定义；完整来源见本页备注

KDRL： 正确性目标与教师约束

历史背景： 2025 年 6 月工作，早于本报告近一年窗口；后续方法的基础对照

$$J_{KDRL} = J_{GRPO} - \beta K_{k2}(p_\theta \| q)$$

$$K_{k2} = E_{a\sim p_{old}}\left[\tfrac{1}{2} (\log p_\theta(a|h) - \log q(a|h))^2 \right]$$

正确性与密集监督

GRPO 优化验证结果
教师 k2 项独立加入目标

Reference 是谁?

KDRL 去掉原始 p_0 的 KL
另加到教师 q 的约束

β 随训练减弱

降低对教师的持续束缚
仍需检查长度与模仿强度

GRPO	KD-RKL	KDRL	KDRL + annealing
54.6	56.1	56.8	**57.2**

7B 教师 → 1.5B 学生； 280 steps ， G=16 ， 20K ；六项宏平均，具体评测口径见备注。

KDRL · arXiv:2506.02208v1 · Eqs.(6)(8), Table 1 ； k2 仅在 on-policy 采样点有相应局部梯度解释

Distilled RL： 细分正优势回答的更新

保留结果优势的符号，用 teacher/student 比值重新分配 token 权重

$$\bar{\rho}_t = \mathrm{clip}(q(a_t|h_t) / p_k(a_t|h_t), 1/3, 3)$$

$$\tilde{\rho}_t = \bar{\rho}_t / \exp(\mathrm{mean}_s \log \bar{\rho}_s)；\quad A^{\mathrm{eff}}_t = w_t A$$

A > 0： $w_t = \tilde{\rho}_t$

在正优势回答中细分学习强度
归一化使权重几何均值为 1

A ≤ 0： $w_t = 1$

保留原始 RL 的负向更新
避免更重地惩罚教师偏好的 token

OPD	GRPO	OPD+RL	Distilled RL
55.97	57.40	56.38	**58.96**

8B-GRPO → Qwen3-4B； 160 steps， G=8， 8K； 10 项数学平均 Pass@1 （32 次采样估计）。

Distilled RL · arXiv:2607.17247v1 · §3.2 Eq.(3), Table 2, App.B

OPDVR： 按结果门控蒸馏信号

先让验证器决定允许的符号，再保留教师给出的 log-ratio 强度

$$r_t = \log(q(a_t|h_t) / p(a_t|h_t))\ ;\quad c = 2R - 1 \in \{-1, +1\}$$

$$\tilde{r}_t = c\ \mathrm{ReLU}(c\ r_t)$$

最终结果	普通 OPD r_t	OPDVR	含义
正确	+0.8	+0.8	认可同一选择：强化
正确	−0.8	0	不压制已有正确路径
错误	+0.8	0	不强化失败路径
错误	−0.8	−0.8	共同否定：压低

$$L = -\Sigma_t\ \mathrm{sg}(\tilde{r}_t)\ \log\ p\theta(a_t|h_t)$$

OPDVR · arXiv:2608.24696v4 · §4.2–4.3； ±0.8 为教学数值， sg 为停止梯度

OPDVR 的证据与适用边界

六基准 Avg@16；两种蒸馏配置分别比较，不能跨行比较模型能力

	普通 OPD	OPDVR	Teacher
Qwen3-4B ← 4B-RL	47.8	49.1	50.4
Qwen3-1.7B-Base ← 4B-Base-RL	20.9	22.8	30.9

六个数学基准 Avg@16， 8K 输出。两行使用不同数据与配置，不能跨行比较模型能力。

$$\text{GRPD：}\ \tilde{r}_{i,t} = \text{sign}(A_i)\ \text{ReLU}(\text{sign}(A_i)\ r_{i,t})$$

sign(A) 决定允许的符号；没有再乘 |A|。当 A = 0，该项也为 0。

门控增加结果验证；论文没有给出公平的端到端墙钟 / 显存对照。

OPDVR · arXiv:2608.24696v4 · Tables 1–2, §4.4； Table 3 使用不同训练数据

OPRD： 用教师方向放大验证梯度

On-Policy Reverse Distillation；适合“较弱教师帮助较强学生”的问题设定

① 教师后训练带来的方向

$$\Delta_t = C(z_T - z_{Tref}) ， \quad d_t = \Delta_t / \|\Delta_t\|_2$$

② 学生自己的结果奖励梯度

$$g_t = A_t \cdot \partial \log p_\theta(a_t|h_t) / \partial z_t$$

各项为什么这样构造

减 reference： 提取教师变化
居中 C： 去掉共同 logit 偏移
单位化 d： 只留下方向
投影：只放大 g

$$\tilde{g}_t = g_t + \lambda_t (d_t^\top g_t) d_t ， \quad \lambda_t \geq 0$$

，g = 0 时无法产生更新。驻点结论限于原文的 logit 。

OPRD · arXiv:2609.08798v1 · §2.2 / App.A； 固定回答、 logit 空间的局部 surrogate

OPRD 的收益、成本与冷启动边界

北京大学 PEKING UNIVERSITY

把学习速度、训练成本与冷启动反例同时看，避免只看最高分

GRPO	OPD	KDRL	OPRD
39.38	39.44	43.99	**51.91**

Qwen3-4B → 8B；四项数学 Mean@16，再跨 step 30/60/90/120/150 平均，非最终分数。

方法额外开销

固定 16K ·4×B200
977.72 → 1094.44 秒 / 步
较 GRPO 增加 11.9%
整体峰值显存 +10.2%

计时的边界

使用交替的人工二元奖励
隔离方法本身开销

未计入真实测试执行
或环境验证成本

冷启动反例

8B-Base → 1.7B-Base
Knights & Knaves 常全失败
OPRD 等仍低于 20%
普通 OPD 达 40.5%

OPRD · arXiv:2609.08798v1 · Table 1, Fig.4(a), App.K；计时实验采用合成奖励

SCOUT： 在教师端使用可验证奖励

北京大学 PEKING UNIVERSITY

同样出现 RLVR ，梯度却落在不同模型上；这是与学生端 reward fusion 的关键区别

学生前缀

学生先生成 $pS = yS_{1:k}$
教师从该前缀续写

教师端 RLVR

验证完整结果
只反传教师生成的续写 token

学生端 OPD

使用更新后的教师评分
学生保持原 OPD 更新

方法	数学平均	代码平均
普通 OPD	49.2	56.6
OPD + 教师从原题做 GRPO	48.8	57.0
SCOUT： 教师从学生前缀做 GRPO	51.4	59.7

4B-Instruct-2507 → 1.7B 。教师每 10 次学生更新训练一次；额外需要教师生成与反向。

SCOUT · arXiv:2609.38360v1 · §3.2, Table 4； 主实验教师每 10 次学生更新训练一次

可验证奖励：机制、条件与成本

相同教师、 prompt 与输出长度下比较；样本数、生成量和真实验证耗时分别记录

方法	验证结果进入哪里	教师相关额外工作	关键失败条件
KDRL	学生 RL 项；可掩码 KD	冻结教师前向	KD 与奖励冲突； β/ 长度敏感
Distilled RL	A 的 token 权重	冻结教师前向	A=0 ； 成功题上教师仍可能误导
OPDVR / GRPD	token 信号的符号门控	冻结教师前向	verifier 错误； GRPD 的零优势组
OPRD	学生 logit 梯度	教师 + teacher reference 前向	无结果梯度； reference 带入偏差
SCOUT	教师续写的 RL 更新	教师生成 + 前向 / 反向	教师不可训练；额外生成 / 验证昂贵

优先诊断：全失败 / ，，。

继续第 07 专题公式选讲（ P64 ）

直接进入内部实验（ P88 ）

机制与成本按原文操作构成归纳，不是跨论文的实测速度排名；完整链接见备注

本节保留完整推导，可按时间与听众背景选讲

Section 07

Objective & exploration

G-OPD

把隐含奖励和 KL 约束拆开，得到可调目标

ExOPD

沿教师相对参考模型的改进方向继续外推

REOPOLD

限制异常负奖励，再按训练阶段选择 token

本节可跳过：点击这里直接进入内部实验（ P88 ）。

[P03] G-OPD / ExOPD §3 · [P06] REOPOLD §3–4

共同符号 | 三种概率与两种 λ

北京大学 PEKING UNIVERSITY

以下先固定一个学生前缀 h，再讨论该前缀上的 token 分布

符号	含义	在反向传播中的角色
pθ(a|h)	当前 Student	需要更新
q(a|h)	冻结 Teacher	提供教师概率
b(a|h)	冻结 reference	G-OPD 的比较基准
pold(a|h)	采集 rollout 的 Student 快照	重要性比的分母
λG / λR	奖励倍率 / 混合系数	来自不同论文，含义与范围不同

本专题统一最大化 J，最小化 L=−J。正优势提高采样 token 的概率。

记号为讲解统一：λG 对应 [P03] λ，λR 对应 [P06] λ；sg 表示停止梯度。

为什么 log 概率比可以当作优势

固定前缀上的 Reverse KL 梯度，可以写成 token reward 加权的策略梯度

目标

$$J_1(\theta) = -D_{\mathrm{KL}}(p_\theta \| q) = \sum_a p_\theta(a) \log \frac{q(a)}{p_\theta(a)}$$

记 R(a)=log q(a)−log pθ(a) 。教师相对更认可该 token 时， R 为正。

求梯度

$$\nabla_\theta J_1 = \mathbb{E}_{a\sim p_\theta}[(R(a)-1)\nabla_\theta \log p_\theta(a)] = \mathbb{E}_{p_\theta}[R(a)\nabla_\theta \log p_\theta(a)]$$

因为 $E_p[\nabla \log p]=\nabla \Sigma_a p(a)=0$ ，常数 − 1 在精确期望下消失。

采样更新

$$L_{\mathrm{local}} = -\mathrm{sg}(R(a)) \log p_\theta(a)$$

sg 让 reward 充当固定权重。等价关系针对局部梯度与匹配的采样分布。

[P03] Eq. (5)–(6) · [P06] §3.1 / App. A.1 · 固定前缀推导为讲解补充。

重要性比 ρ 为什么要放在损失里

北京大学 PEKING UNIVERSITY

ρ 修正旧 Student 采样的动作权重， reward 决定更新的方向和强度

$$\rho_t(\theta) = \frac{p_\theta(a_t \mid h_t)}{p_{\text{old}}(a_t \mid h_t)}$$

$$\mathbb{E}_{a\sim p_\theta}[f(a)] = \mathbb{E}_{a\sim p_{\text{old}}}[\rho(a) f(a)]$$

$$L = -\mathbb{E}_{p_{\text{old}}}[\rho_t(\theta)\,\text{sg}(A_t)]$$

为什么需要

同一批 rollout 做多次更新后，当前 Student 已不同于采样快照。

为什么 ρ=1 仍有梯度

θ=θold 时分母冻结。

此时 ∇ ρ=∇log pθ 。

PPO clip 的作用

抑制过大的更新激励，但不能直接消除 reward 的负向长尾。

适用边界

动作级校正不自动修复旧前缀的状态分布偏移。

[P06] Eq. (1)，App. A.1–A.2 · [P03] 官方实现中的 PPO surrogate 。

G-OPD 的动机｜奖励与约束为何要解耦

标准 OPD 已包含一个稠密奖励，但它与 KL 正则的相对权重固定

$$r^{\mathrm{imp}}(a) = \log \frac{q(a)}{b(a)}$$

$$J_1 = \mathbb{E}_{p_\theta}[r^{\mathrm{imp}}] - D_{\mathrm{KL}}(p_\theta \| b)$$

$$J_{\lambda_G} = \lambda_G \mathbb{E}_{p_\theta}[r^{\mathrm{imp}}] - D_{\mathrm{KL}}(p_\theta \| b)$$

Teacher 提供什么

q/b 描述教师相对 reference 的概率提升，不等同于答案正确率。

标准 OPD 限制在哪里

奖励项与 KL 项绑定为 1:1，理想最优分布停在 Teacher 。

新增哪个自由度

λG 调节相对权重， b 定义“沿哪个模型变化方向”学习。

G-OPD 提供目标族； ExOPD 是其中 λG>1 的选择。

[P03] §3.2，Eq. (7)，(9)，(11)

G-OPD 的分解｜为何加上再减去 log b

引入 reference 后，原来的教师匹配目标可以分成奖励和约束两部分

加减同一项

$$\log \frac{q}{p_\theta} = \underbrace{\log \frac{q}{b}}_{\text{implicit reward}} - \underbrace{\log \frac{p_\theta}{b}}_{\text{sampled KL term}}$$

这一步只是恒等变形，没有增加额外训练信号。

取学生期望

$$\mathbb{E}_{p_\theta}\left[\log \frac{q}{p_\theta}\right] = \mathbb{E}_{p_\theta}\left[\log \frac{q}{b}\right] - D_{\mathrm{KL}}(p_\theta \| b)$$

KL 本身是期望，不能把一个 token 的 log ratio 当成完整 KL。

只缩放奖励

$$J_{\lambda_G} = \mathbb{E}_{p_\theta}[\lambda_G \log q + (1-\lambda_G)\log b - \log p_\theta]$$

λG=1 时 b 完全抵消，恢复标准 OPD； λG≠1 时 b 影响目标。

[P03] Eq. (7)、(11)；将 KL 移到期望外，展开统一记号。

G-OPD 的优势｜三个 log 项各做什么

统一采用正向优势 AG，最小化损失时再加整体负号

$$A_t^G = \lambda_G \log \frac{q_t}{b_t} - \log \frac{p_{k,t}}{b_t}$$

$$A_t^G = \underbrace{\log \frac{q_t}{p_{k,t}}}_{\text{standard OPD}} + (\lambda_G - 1) \underbrace{\log \frac{q_t}{b_t}}_{\text{reference improvement}}$$

$$L_G = -\mathbb{E}_{p_k}\left[\rho_t(\theta)\,\mathrm{sg}(A_t^G)\right]$$

教师项 λG log q

推动学生学习教师偏好的动作。

参考项 (1−λG) log b

λG>1 时提高相对 reference 的改进权重。

学生项 − log pk

提供分布约束，使已被学生过度赋权的 token 降温。

实现中的符号

官方代码用 advantages=−reverse_kl，与此处 AG 一致。

ExOPD 额外加的是 (λG−1)log(q/b)，并非把整个 OPD loss 乘 λG。

[P03] Eq. (14) / App. A；统一正向优势与 γ=0 局部近似；官方实现 RUCBM/G-OPD。

最优分布｜归一化常数从哪里来

固定 h，允许 p 在概率单纯形上自由变化，并假设 q、b 为正

约束优化

$$\mathcal{L}(p,\nu)=\sum_a p(a)\left[\lambda_G \log\frac{q(a)}{b(a)}-\log\frac{p(a)}{b(a)}\right]+\nu\left(\sum_a p(a)-1\right)$$

最后一项保证所有 token 概率之和为 1。

一阶条件

$$\frac{\partial\mathcal{L}}{\partial p(a)}=0 \quad\Longrightarrow\quad \log p(a)=\lambda_G\log q(a)+(1-\lambda_G)\log b(a)+\nu-1$$

ν−1 对所有 token 相同，正好承担归一化作用。

闭式解

$$q^*_{\lambda_G}(a)=\frac{q(a)^{\lambda_G}b(a)^{1-\lambda_G}}{Z_{\lambda_G}},\qquad Z_{\lambda_G}=\sum_v q(v)^{\lambda_G}b(v)^{1-\lambda_G}$$

因此 JλG=log ZλG−KL(p‖q*λG)，最优点为该归一化分布。

[P03] Eq. (12) 的补充推导，显式补上归一化；结论限于固定状态无约束分布。

ExOPD 的外推｜放大哪一段 odds 变化

外推作用在教师相对 reference 的 log-odds 增量上

$$\frac{q^*_{\lambda_G}(a)}{q^*_{\lambda_G}(c)} = \frac{b(a)}{b(c)}\left[\frac{q(a)/q(c)}{b(a)/b(c)}\right]^{\lambda_G}$$

$$\log\frac{q^*_{\lambda_G}(a)}{q^*_{\lambda_G}(c)} = \log\frac{b(a)}{b(c)} + \lambda_G\left[\log\frac{q(a)}{q(c)} - \log\frac{b(a)}{b(c)}\right]$$

λG=0	**停在 reference b**
0<λG<1	**在 b 与 q 的 log-odds 之间插值**
λG=1	**恢复教师分布 q**
λG>1	**沿同一变化方向超过教师的 log-odds**

“超过教师”指目标分布外推。任务准确率能否超过教师仍需实验。

[P03] §3.2； odds 形式为对归一化最优解的讲解推导。

三 token 例子｜外推不等于强化教师第一名

教学示例： reference b=(0.70,0.20,0.10)， Teacher q=(0.60,0.30,0.10)

token	reference b	Teacher q	q/b	ExOPD 目标 λG=2
A	0.700	0.600	0.857	0.483
B	0.200	0.300	1.500	0.423
C	0.100	0.100	1.000	0.094

$$q_2^* \propto q^2/b = (0.5143,\ 0.4500,\ 0.1000), \qquad Z_2 = 1.0643$$

A 虽然仍是 Teacher 概率最高的 token，但 Teacher 相比 reference 降低了它。

B 的相对提升最大，因此外推后 B 从 0.300 增至 0.423。 λG=2 仅用于演示。

根据 [P03] 的归一化最优解自行计算；本页数值不是论文实验。

奖励倍率、学习率与温度的区别

判断一个改动是在改变优化目标，还是只改变抵达目标的过程

操作	改变了什么	理想最优分布
增大学习率 η	同一梯度方向的步幅	目标不变；过大可能不稳定
整体乘 $c \cdot J_1$ ， $c>0$	所有目标项同比缩放	仍为 q
G-OPD 改 λG	奖励与 KL 的相对权重	$q^*_{\lambda_G} \propto q^{\lambda_G} b^{1-\lambda_G}$
Teacher 温度 T	对教师 logits 作统一缩放	$q_T \propto q^{1/T}$

b 为均匀分布时，目标退化为教师温度 T=1/λG 的变换；一般情形不等价。

默认 b 为初始 Student；若能得到 Teacher 的前身 checkpoint，也可将其作为参考方向。

[P03] §3.2，reference 消融；学习率 / 温度对照为讲解分析。

G-OPD 与 ExOPD | 实验支持到哪里

北京大学 PEKING UNIVERSITY

同论文、同一 Qwen3-4B 多教师整合配置；对比 λG=1 与 1.25

平均准确率 / %	领域 Teacher	OPD λG=1	ExOPD λG=1.25	相对 OPD
数学四集 Avg@32	46.0	46.4	47.7	+1.3 pp
代码三集 Avg@4	61.2	60.6	62.0	+1.4 pp

强到弱设置：相对 OPD 提升，仍低于 Teacher

强到弱数学均值	OPD	ExOPD	30B-A3B Teacher
Student 1.7B	23.1	25.4	59.7
Student 4B	42.6	45.3	59.7

同源多教师： reference 为初始 Qwen3-4B-Non-Thinking， Teacher 是其 Math / Code RL 变体。
强到弱时优于 OPD，但未超过教师；较大 λG 可能放大偏差，生成长度也可能增长。

[P03] Tables 2–3 · Figure 4；完整配置与生成长度限制见讲稿。

REOPOLD 的动机｜失稳来自哪些 token

把教师的 log-prob ratio 当作 reward 后，可以直接诊断更新信号

$$R_t = \log \frac{q_t}{p_{\theta,t}}$$

$$\frac{q_t}{p_{\theta,t}} \to 0 \implies R_t \to -\infty$$

$$\frac{q_t}{p_{\theta,t}} \approx 1 \implies R_t \approx 0$$

负奖励长尾

少数极端负值可能占据很大的梯度权重，强力压低学生动作。

接近零的奖励

许多已对齐 token 提供很小的直接学习信号。

探索过早收缩

早期强负反馈可能使学生很快集中于少数路径。

对应设计

负值截断、按熵筛选、早期与后期使用不同 mask 。

这些是论文观察到的优化问题，并非“所有负奖励都没有价值”。

[P06] §3.2 · Figures 4–6

REOPOLD 的混合分布｜为什么掺入 Student

教师几乎不给概率的动作，仍能从学生分布获得有限的概率质量

构造混合

$$m_{\lambda_R} = (1-\lambda_R)q + \lambda_R p_\theta, \qquad 0 < \lambda_R < 1$$

q 很小时， m 仍保留 λR pθ ，避免混合分布相对学生的概率比趋于 0 。

Jensen 不等式

$$\log m_{\lambda_R} \geq (1-\lambda_R)\log q + \lambda_R \log p_\theta$$

因为 log 是凹函数。两边减去 log pθ ，再除以正数 1−λR 。

得到上界

$$R = \log \frac{q}{p_\theta} \leq U_{\lambda_R}(R) = \frac{1}{1-\lambda_R} \log\left[(1-\lambda_R)e^R + \lambda_R\right]$$

混合式 U 是原始 reward 的上界；下一页取它在负向尾部的极限。

[P06] §4.1； App. A.2 Eq. (20)–(21)

负奖励下界｜分母 1−λR 从哪里来

分母来自 Jensen 不等式的重排，log 内只有 λR

负向尾部极限

$$\lim_{R\to-\infty} U_{\lambda_R}(R) = \frac{\log\lambda_R}{1-\lambda_R} \equiv b_R$$

此时 e^R 趋于 0，混合比值只剩 λR。

裁剪规则

$$\widehat{R}_t = \max\left(\mathrm{sg}(R_t), \frac{\log\lambda_R}{1-\lambda_R}\right)$$

作者取这个有限极限作为裁剪阈值，只修正过度负向的 reward。

数值和边界

$$\lambda_R = 0.3: \ b_R \approx -1.720; \qquad \lim_{\lambda_R\to 0^+} b_R = -\infty, \qquad \lim_{\lambda_R\to 1^-} b_R = -1$$

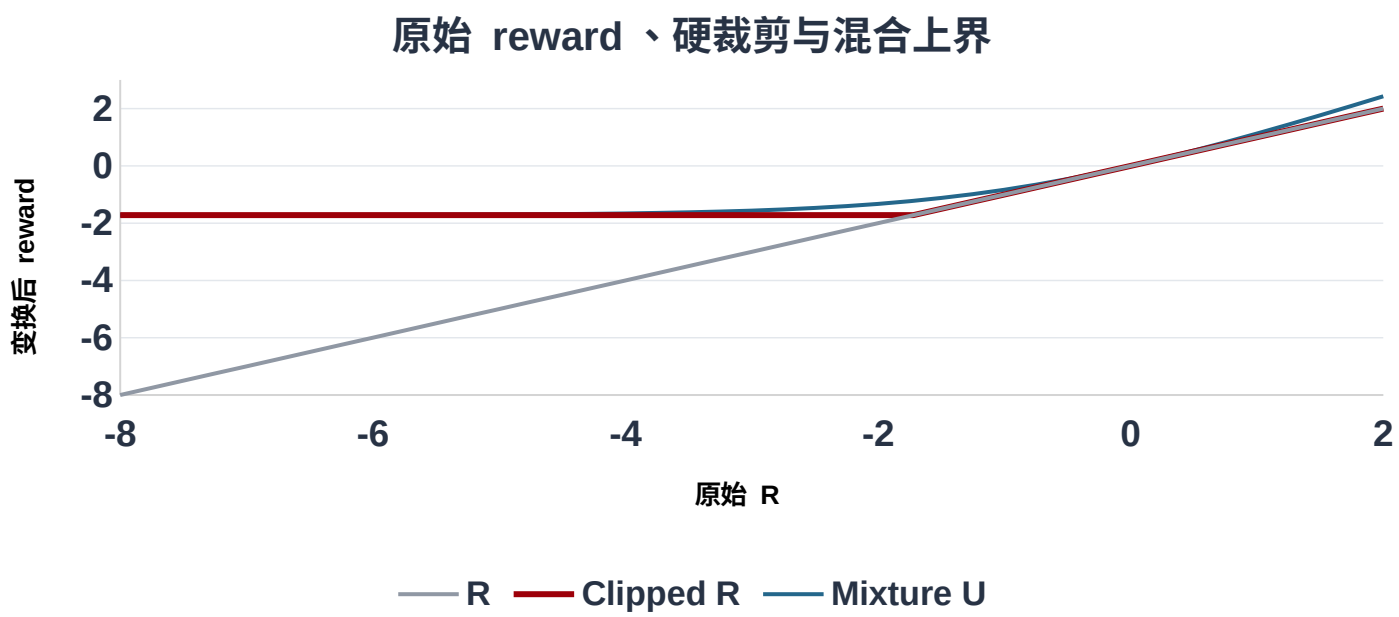
λR 越大，下界越高，负惩罚限制越强；论文使用固定 λR=0.3。

[P06] Eq. (5)、(7)、App. A.2；边界极限为讲解推导。

裁剪的形状｜只改变负向尾部

λR=0.3 时，低于 − 1.720 的奖励抬到同一下界

$$\widehat{R} = \max(R, -1.720), \qquad R \le \widehat{R} \le U_{0.3}(R)$$



原始 R	裁剪后 R̂
−8.000	−1.720
−4.000	−1.720
−1.000	−1.000
+1.000	+1.000

正奖励和温和负奖励保持原值。

混合曲线 U 与 max 裁剪不是同一个函数。

推导给出阈值依据，并未证明“裁剪梯度等于混合 KL 梯度”。

按 [P06] Eq. (7)、(21) 绘制的函数图；数值为讲解计算。

Reward clipping 与 token masking 的区别

北京大学 PEKING UNIVERSITY

clip 改变权重， mask 决定这个 token 是否直接进入损失

同一个 token	保留的 reward	直接损失贡献
原始 R=−8	−8	可能形成很强负更新
只做 clipping	−1.720	仍有有限负更新
早期 reward mask	M=0	该 token 的直接损失项为 0
后期 entropy mask，且高熵	M=1， $\hat{R}$=−1.720	以有限负权重重新进入更新

$$M_t^{\text{early}} = \mathbb{I}[R_t \geq b_R], \qquad M_t^{\text{late}} = \mathbb{I}[H_t \geq \tau_\beta]$$

早期保留的 token 都满足 R≥bR，因此这些 token 的 clipping 不改变数值。

M=0 只表示直接 loss 被屏蔽，共享参数更新仍可能改变该 token 概率。

[P06] Eq. (5)–(6)，Algorithm 1；相同 token 的对照为讲解示例。

Student entropy ｜为什么保留最高 β 比例

熵用于寻找学生仍不确定的位置；它不是教师置信度，也不是正确率

$$H_t = -\sum_{v \in \mathcal{V}} p_\theta(v \mid h_t) \log p_\theta(v \mid h_t)$$

$$\tau_\beta = \mathrm{Quantile}_{1-\beta}(\{H_t\}_{\mathcal{B}})$$

$$M_t^{\mathrm{late}} = \mathbb{I}[H_t \geq \tau_\beta]$$

经验依据

论文中低熵位置的 reward 大量聚集在 0 附近，直接学习信号弱。

β=0.2 的含义

保留 batch 中预测熵最高的约 20%，阈值在 80 分位。

选择高熵的目的

把直接更新集中到学生更不确定的分支位置。

与 EOPD 的差别

REOPOLD 用 Student 熵选 token； EOPD 用 Teacher 熵决定补 FKL。

[P06] §4.2，Eq. (8)，PDF Figures 5，8；与 [P05] 的比较为报告整理。

阶段调度｜真正切换的是 mask

λR 保持固定，早期保护探索，后期按 Student 熵筛选

$$M_t^{(k)} = \begin{cases} \mathbb{I}[R_t \geq b_R], & k < T_{\text{switch}}, \\ \mathbb{I}[H_t \geq \tau_\beta], & k \geq T_{\text{switch}}. \end{cases}$$

探索阶段 屏蔽极强负反馈，温和负奖励和正奖励仍可学习。

精修阶段 使用熵 mask 替换 reward mask，高熵的极负 token 也可经 clipping 参与。

示例配方 数学 1.5B： 300 steps，在 100 steps 切换； λR=0.3 固定。

公式没有把两个 mask 相乘，也没有要求随阶段增大 λR。

[P06] §4.3，Eq. (6)，Algorithm 1； PDF Table 8。

完整目标｜每个因子分别解决什么

选中 token 的平均收益，把信号大小、采样校正和选择机制放在同一式中

$$J_{\text{REOPOLD}} = \mathbb{E}\left[\frac{\sum_{i,t} M_{i,t}^{(k)}\, \rho_{i,t}(\theta)\, \widehat{R}_{i,t}}{\sum_{i,t} M_{i,t}^{(k)}}\right], \qquad L = -J_{\text{REOPOLD}}$$

ρi,t	旧 Student 采样下的动作权重校正
R̂i,t	经停止梯度和负值截断后的更新信号
Mi,t 与分母	只对选中 token 计直接损失，并按选中数归一化
计算成本	loss mask 不会自动省去完整 rollout、Teacher scoring 或 Student forward

保留同一组 20% token 时，若分母仍用总 token 数，梯度缩为原来的约 1/5 。

[P06] Eq. (4)、(8)；归一化与实际计算成本解释为讲解分析。

REOPOLD | 同训练轮数结果与消融

同 Teacher、同 Student、同 rollout steps；这不等于同 FLOPs 或墙钟时间

数学能力迁移

数学 1.5B / 300 rollout steps	AIME24 Avg@32	AIME25 Avg@32
Vanilla RKL	37.1	30.6
REOPOLD	41.6	32.6
绝对变化	+4.5 pp	+2.0 pp

左： SkyWork-OR1-Math-7B → DeepSeek-R1-Distill-Qwen-1.5B 。

300 rollout steps ，每轮还包含 minibatch 更新。 Avg@32 表示平均准确率。

[P06] PDF Table 1、Table 6；配方及指标口径见讲稿。

视觉 3B 模块累加消融

视觉 3B 的累加消融	均值 / %
RKL	35.37
+ stop-gradient	39.33
+ reward clipping	39.73
+ entropy sampling	40.81
+ stage schedule	41.12

右： Qwen2.5-VL-3B ， 60 rollout steps 。各模块按顺序累加，不能视为独立贡献。

三种方法｜它们改变的数学对象不同

北京大学 PEKING UNIVERSITY

G-OPD 是目标族， ExOPD 是外推区域， REOPOLD 调整奖励与 token 选择

维度	G-OPD	ExOPD	REOPOLD
主要动机	解耦奖励与 KL	沿改进方向超出教师	缓解极负反馈和无效更新
核心改动	λG 与 reference b	选择 λG>1	R̂ + 阶段性 mask
Teacher 信息	采样 token 的 q	采样 token 的 q	采样 token 的 q
额外信息	reference 概率 b	reference 概率 b	Student 预测熵
主要风险	参考模型选错方向	放大偏差、输出变长	筛选偏置、有效样本变少

二者可以提出组合实验，但稳定性和收益不能从单篇结果直接推出。

综合 [P03] §3–4 与 [P06] §3–5；比较及组合判断为报告分析。

推导边界｜局部目标与完整序列要区分

局部公式便于解释和实现，但不能直接当作完整序列优化的无条件证明

$$G_t = \sum_{u=t}^{T} R_u, \qquad \nabla J_{\text{seq}} \text{ contains } \mathbb{E}[G_t \nabla \log p_\theta(a_t \mid h_t)]$$

$$\gamma = 0: \quad G_t \approx R_t$$

固定前缀的结论

本专题最优 q* 与零均值 score 恒等式，在明确的条件下成立。

序列还会改变后续状态

当前 token 影响未来前缀，完整回报包含后续位置的奖励。

神经网络还受容量约束

不同前缀共享参数，未必能逐状态精确达到闭式最优分布。

实现核对

明确采样温度、 old snapshot 、reward 、mask 分母和 loss 符号。

[P03] Eq. (5)–(6) 、App. A； 其余为实现与条件分析。

本专题小结｜公式设计对应的四个问题

北京大学 PEKING UNIVERSITY

把论文中的每一项放回它要解决的优化问题

为什么引入 reference	把“教师相对改进”从“学生偏离约束”中分离出来。
为什么 λG>1 称为外推	放大的是相对 reference 的 log-odds 增量，最优目标发生改变。
为什么使用这个负奖励下界	Jensen 混合上界在教师相对概率趋零时，收敛到这个有限值。
为何分阶段改变 mask	先屏蔽强负反馈保护探索，再把直接学习集中于高熵位置。

比较效果时，同时报告准确率、多样性、生成长度和实际训练成本。

[P03] G-OPD / ExOPD · [P06] REOPOLD；推导、例子和方法边界见对应讲稿。

北京大学 PEKING UNIVERSITY

把理论假设落到统一预算、学习曲线和多种子实验

Section 08

Our Evidence

01 10-update OPD：早期信号

02 性能与成本：不同计量口径

03 研究计划：自适应刷新机制

本节结论：先检验机制，再讨论论文级提升；保留显著性与预算约束。

本组实习自有实验汇总（MOPD： 3 seeds； OPD： 10-update），不同于公开论文结果

怎样选择方法：先定位当前瓶颈

北京大学 PEKING UNIVERSITY

不同论文的学生、任务和计算口径不同；下面按问题选择起始对照

观察到的问题	优先对照	必须同时记录
轨迹生成占据主要成本	全长 OPD / Prefix / Semi	总 GPUh、长度上限、精度
多样性或 Pass@k 下降	OPD / EOPD / REOPOLD	Avg@n、Pass@n、长度、熵
教师在学生前缀上失效	OPD / SCOUT / ESR	续写恢复率、教师训练预算
短领域任务被长任务淹没	Naive M-OPD / Open-MOPD	token 份额、每域进度
只有外部教师文本	SFT / GAD / BSD	、、

最值得先测的三项：师生兼容性、有效 token 预算、端到端成本

[P04] Prefix OPD [P05] EOPD [P11] Open-MOPD [P13] SCOUT [P15] Semi-OPD [P18] BSD

内部复现： OPD 统一的早期对比

北京大学 PEKING UNIVERSITY

给定共同学生 / 教师锚点；训练仅 10 次更新；采用同一最终评测

Initial student

31.45%

共同初始准确率

Teacher

62.60%

上界参考，非同预算比较

Final candidates

20 行

16 项标记 formal 的方法范围

评测： 256 道题 × 每题 4 次回答

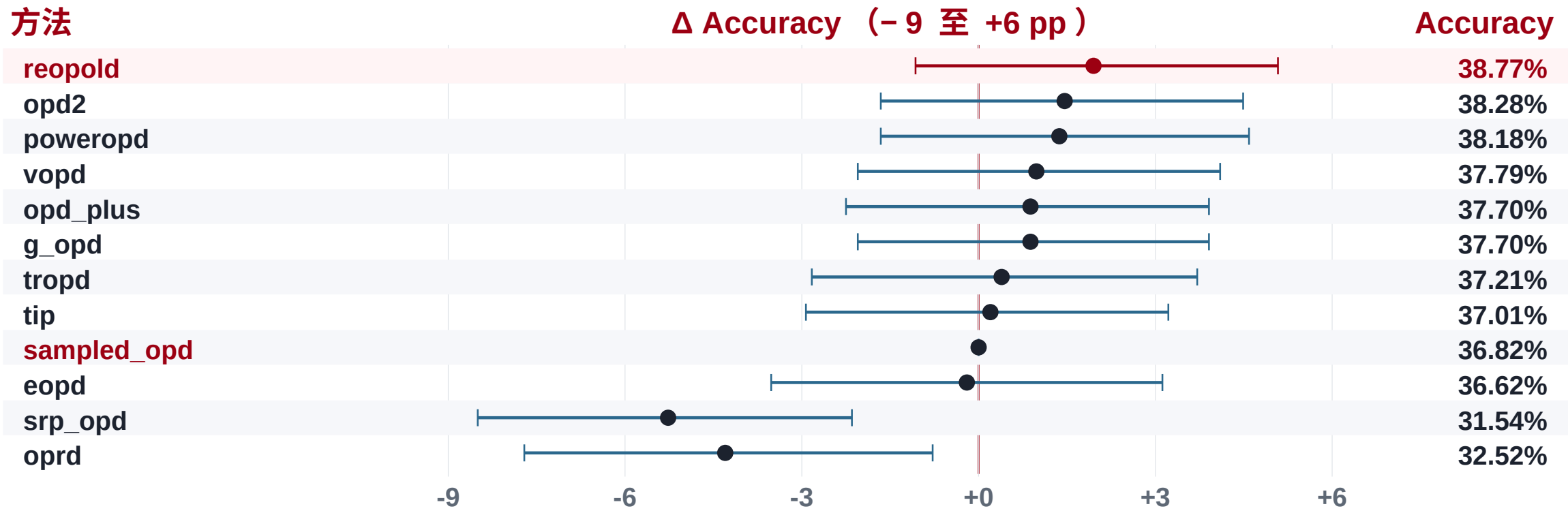
区间：题级配对 percentile bootstrap ， 95% ；条件于一次训练过程

这批 10-update 对比只观察训练初期：不代表收敛性能或全任务总体优势。

来源：作者提供的实习实验自动汇总；内部结果，非原论文报告

OPD · 十次更新后的准确率差异

相对 sampled OPD 的配对差值与 95% bootstrap 区间（单位：百分点）

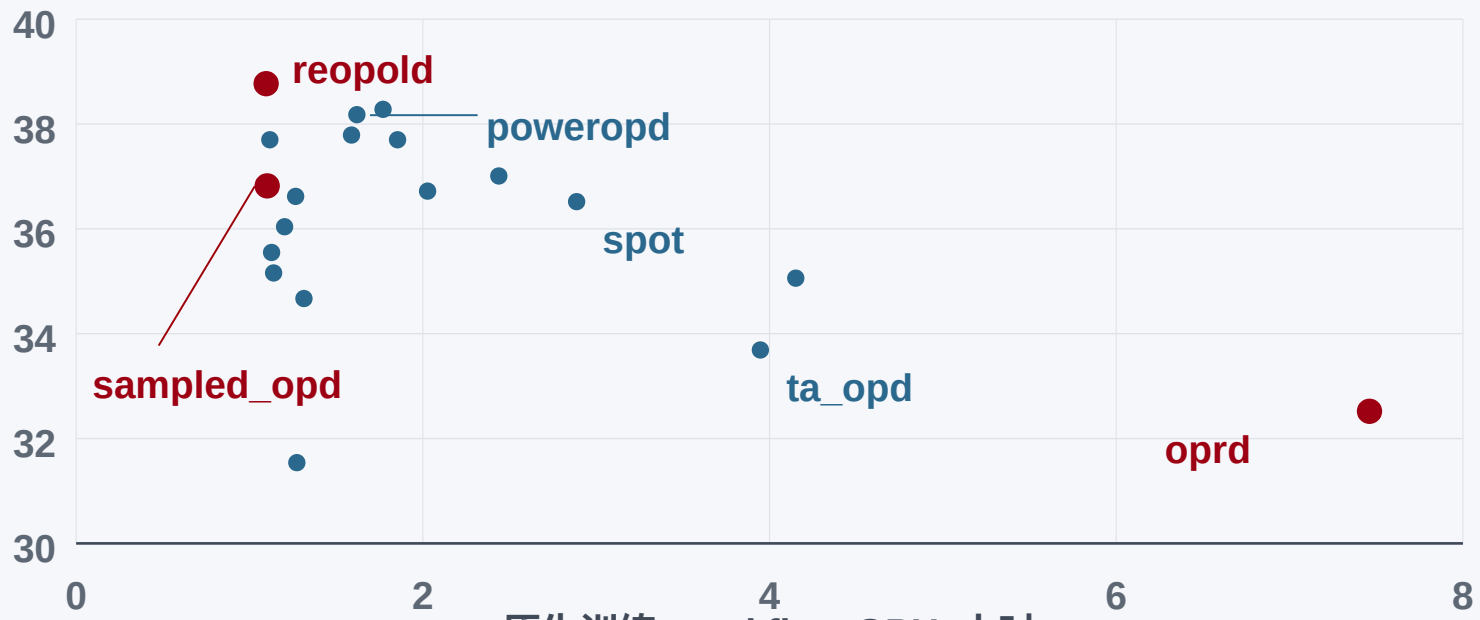


REOPOLD +1.95 pp 的区间 [−1.07, +5.08] 包含 0；本次不能宣称显著优于 sampled OPD 。

来源：作者提供的实习实验自动汇总；内部结果，非原论文报告

OPD · 早期收益与原生 workflow 成本

描述性成本—性能散点；框架和 GPU/ 角色分配不同，不作严格公平 Pareto 排名



关键对照

REOPOLD
38.77% · 1.096 h

Sampled OPD
36.82% · 1.102 h

PowerOPD
38.18% · 1.619 h

OPRD
32.52% · 7.461 h

TrOPD 的 teacher-bank 准备另需 0.376 GPUh；缺失的原生成本不以 0 代替。

来源：作者提供的实习实验自动汇总；内部结果，非原论文报告

OPD · 十次更新告诉了我们什么?

北京大学 PEKING UNIVERSITY

偏早期可训练性、损失稳定性和计算开销诊断；并非最终方法排名

01 优势方法仍存在区间重叠

REOPOLD ，OPD2 ，PowerOPD 对 sampled 的差值区间均包含 0 。

02 早期训练可能明显不稳定

SRP-OPD +0.10pp 对学生几乎无提升； OPRD 较 sampled 低 4.30pp 。

03 性能和效率要拆开复现

成本包含框架原生流程，但 GPU/ ，。

04 后续应扩展到完整收敛

每法多种子、统一训练 token/GPUh ，完整曲线与多重比较控制。

与前面 MOPD 三种子实验不能合并估计误差： OPD 区间只条件于单次训练。

来源：作者提供的实习实验自动汇总；内部结果，非原论文报告

可研究的问题：何时值得刷新轨迹?

内部早期实测提供线索；以下三条具体机制假设仍需独立验证

假设 H1：初始兼容性决定复用收益的方向

用相近能力、不同兼容性的师生对，分离“教师强”与“教师适配”

假设 H2：训练中的分布漂移决定刷新时机

比较固定数据、固定周期刷新与基于漂移的部分刷新

假设 H3：局部纠正价值决定刷新哪些样本

在相同新增 token 预算下，检验均匀抽样与价值抽样

证伪标准：相同总 GPUh 下，简单的周期刷新能否达到同样结果?

[P09] ESR [P13] SCOUT [P15] Semi-OPD [P16] Demystifying OPD

实验设计：把分布收益与算力收益拆开

内部已有小规模验证；下一阶段需对齐训练预算并扩展到完整收敛

实验要素	建议设置
师生组合	实测高 / 低兼容性；覆盖数学与代码
方法对照	SFT、OPD、Semi、周期刷新、部分自适应刷新；另加 Prefix
统一条件	、、cap、采样温度、checkpoint 选择规则
性能指标	Avg@n / Pass@n、域外任务、长度、每域退化
预算与诊断	总 GPUh + 教师调用量； overlap、续写恢复率、缓存漂移
稳定性	预先固定评测集；建议 ≥ 3 个训练 seed；报告各次结果

[P04] Prefix OPD [P13] SCOUT [P15] Semi-OPD

总结与讨论

最新论文与内部实验共同提示：性能、稳定性和成本都必须控制变量

轨迹新鲜度是一项可以测量的取舍

在线访问覆盖、教师兼容性与生成成本，需要共同考虑

教师反馈的质量是条件化的

原题能力、学生前缀恢复能力、局部梯度可教性彼此不同

目标函数决定哪些能力容易迁移

‘Avg@n 与 Pass@n ’领域均衡应分别检验

讨论：哪些题目最值得蒸馏？何时需要验证奖励？何时刷新轨迹？

[P05] EOPD [P11] Open-MOPD [P13] SCOUT [P14] Skills / Knowledge [P15] Semi-OPD

2026 年 9–10 月：新的问题与证据

北京大学 PEKING UNIVERSITY

近预印本按首次公开日期列示；机制与结论边界见正文相关专题

首发日期	工作	本页之后要回答的问题
09.23	RT-OPD	如何让蒸馏更依赖音频证据?
09.29	SCOUT	教师是否也需要适应学生前缀?
10.07	Skills / Knowledge	OPD 迁移了新事实，还是组合能力?
10.08	Semi-OPD	持续刷新学生轨迹何时有益?
10.13	Demystifying OPD	教师的有效纠正发生在哪里?
10.15	TF-OPD	同源专家能否用闭式方法整合?
10.19	BSD	只有教师文本，如何构造 token 信号?

[P12] RT-OPD [P13] SCOUT [P15] Semi-OPD [P18] BSD

序列 RKL 梯度与局部 surrogate

固定教师 q； $h_t = (x, y_1, \ldots, y_{t-1})$；以下由序列 KL 定义直接推导

精确序列目标

$$\mathcal{L}_{\text{seq}}(\theta) = \mathbb{E}_{y \sim \pi_\theta}\left[\log \frac{\pi_\theta(y \mid x)}{q(y \mid x)}\right]$$

$$A_{\theta,t} = \log q(y_t \mid h_t) - \log \pi_\theta(y_t \mid h_t), \qquad G_{\theta,t} = \sum_{u=t}^{T} A_{\theta,u}$$

$$\nabla_\theta \mathcal{L}_{\text{seq}} = -\mathbb{E}_{y \sim \pi_\theta} \sum_t G_{\theta,t} \nabla_\theta \log \pi_\theta(y_t \mid h_t)$$

局部 surrogate：只保留当前位置的 A

$$g_{\text{local}} = -\mathbb{E}_{y \sim \pi_k} \sum_t \text{sg}(A_{k,t}) \nabla_\theta \log \pi_\theta(y_t \mid h_t)|_{\theta=\theta_k}$$

局部等式前提：固定前缀分布，采样策略与学生 log-prob 对应，教师在同一前缀评分。

关键区别：局部 surrogate 忽略后续状态访问效应，≠ 完整序列 RKL 梯度。

[B1] MiniLLM §2 / 序列目标　[B2] GKD §3　[T1] Thinking Machines Implementation

补充方法： TrOPD 与 Direct-OPD

在各自固定师生与训练设置内比较；结果据原论文表格重绘

TrOPD： 按可信区域选择监督

Qwen3-Nemotron-4B → Qwen3-SFT-1.7B

可信 token 用 RKL；离群 token 用 top-k FKL。
教师前缀指导逐步退火至零。

同设置 / 多领域	均值（%）
OPD	48.29
TrOPD	51.73

同论文均值增加 3.44 pp

[P19] TrOPD Table 4 / §3

Direct-OPD： 迁移教师 RL 增量

JustRL-1.5B 教师对 → Qwen3-4B 学生

r = log qpost-RL − log qpre-RL

以强学生初始策略为 KL 参考；需两个教师 checkpoint。

AIME / Avg@32	2024	2025
弱教师	51.3	37.5
初始学生	72.5	65.6
Direct-OPD	77.6	68.8

相对初始学生 +5.1 / +3.2 pp

[P20] Direct-OPD Table 1 / §3

比较 OPD 结果前，先对齐统计口径

北京大学 PEKING UNIVERSITY

分数、计算与不确定性对应不同对象；先核对定义，再解释差值

口径	它测量什么	报告与对齐要求
Avg@n	每题 n 次采样的平均正确率	固定 n、题集、提示模板与采样温度
Pass@n	n 次内至少一次成功的题目比例	固定 n；核对是否采用 pass@k
pp / 相对提升	pp = 新值−旧值； 相对提升 = (新值−旧值)/ 旧值	百分比指标的差用 pp； 避免把 relative % 写成百分点
计算成本	GPUh：各 GPU 小时总和； wall-clock：经过时间； FLOP：运算量	说明 GPU 型号、并行度与计时范围； 是否含教师训练、采样、评分和评测
Headroom	(分数−起点)/(参考上限−起点)	明确起点、参考上限与跨域聚合方式； 不同参考下的恢复率不可直接排名
种子与协议	训练 seed、推理 seed； 训练与测试长度上限	、、、tokenizer、 / top-p 与 checkpoint 选择

[P05] EOPD Table 2　[P10] MOPD Table 2　[P11] Open-MOPD Table 2　[P15] Semi-OPD Table 4　[P22] ActFirst-OPD §4

附表： MOPD 设定 A · 全部方法

北京大学 PEKING UNIVERSITY

三种子结果按给定“恢复率”列求均值 ± 样本 SD；步时为每个 seed 中位步时的均值

方法	Math / %	Code / %	IF / %	中位步时 / s
param_avg	50	29	31	—
param_ta	137	148	129	—
dn_mopd	106 ± 10	97 ± 16	93 ± 25	620
mopd_pg	-365 ± 0	-375 ± 0	-885 ± 3	522
mopd_topk	112 ± 9	99 ± 8	117 ± 12	556
open_mopd	115 ± 12	116 ± 16	118 ± 44	930
road_mopd	99 ± 10	92 ± 4	119 ± 20	690
single_code	12 ± 11	106 ± 11	-16 ± 45	557
single_if	-22 ± 10	-4 ± 10	99 ± 34	557
single_math	84 ± 16	34 ± 6	-12 ± 33	627
trust_mopd	110 ± 13	85 ± 19	123 ± 26	638
uniform_opd	62 ± 20	62 ± 20	84 ± 17	618
up_mopd	124 ± 22	102 ± 5	119 ± 14	620

PG 明显崩溃；参数基线仅单次结果、无方差。绝对 benchmark 分数见原始记录。

来源：作者提供的实习实验自动汇总；内部结果，非原论文报告

附表： MOPD 设定 B · 全部方法

北京大学 PEKING UNIVERSITY

三种子结果按给定“恢复率”列求均值 ± 样本 SD；只有数学和代码两领域

方法	Math / %	Code / %	中位步时 / s
param_avg	85	107	—
param_ta	80	43	—
dn_mopd	64 ± 4	83 ± 3	113
mopd_pg	-46 ± 3	-36 ± 6	73
mopd_topk	65 ± 4	77 ± 4	112
open_mopd	67 ± 8	71 ± 5	320
road_mopd	68 ± 4	82 ± 7	140
single_code	36 ± 3	86 ± 3	113
single_math	74 ± 2	41 ± 10	112
trust_mopd	60 ± 6	70 ± 4	125
uniform_opd	66 ± 4	76 ± 8	112
up_mopd	71 ± 2	86 ± 13	115

B 不评测 IF； param_avg 的 85% / 107% 是当前必须对齐的强参考线。

来源：作者提供的实习实验自动汇总；内部结果，非原论文报告

附表： OPD · 全部 20 行方法 / 变体

北京大学 PEKING UNIVERSITY

最终 checkpoint ｜ 相对 sampled OPD 的 95% 配对 bootstrap 区间 ｜ native workflow GPUh

方法	Acc/%	Δ [95% CI] / pp	GPUh
reopold	38.77	+1.95 [-1.1,+5.1]	1.10
opd2	38.28	+1.46 [-1.7,+4.5]	1.77
poweropd	38.18	+1.37 [-1.7,+4.6]	1.62
vopd	37.79	+0.98 [-2.0,+4.1]	1.59
opd_plus	37.70	+0.88 [-2.2,+3.9]	1.12
g_opd	37.70	+0.88 [-2.0,+3.9]	1.85
tropd	37.21	+0.39 [-2.8,+3.7]	—
tip	37.01	+0.20 [-2.9,+3.2]	2.44
sampled_opd	36.82	+0.00 [+0.0,+0.0]	1.10
scope	36.72	-0.10 [-3.0,+3.0]	2.03

方法	Acc/%	Δ [95% CI] / pp	GPUh
eopd	36.62	-0.20 [-3.5,+3.1]	1.27
spot	36.52	-0.29 [-3.5,+2.8]	2.89
topk16_opd	36.04	-0.78 [-4.0,+2.3]	1.20
opdvr	35.55	-1.27 [-4.3,+1.9]	1.13
eos_corrected_opd	35.16	-1.66 [-4.7,+1.5]	1.14
aopd	35.06	-1.76 [-5.2,+1.7]	4.15
local_lp_opd	34.67	-2.15 [-5.4,+1.2]	1.31
ta_opd	33.69	-3.12 [-6.2,+0.0]	3.95
oprd	32.52	-4.30 [-7.7,-0.8]	7.46
srp_opd	31.54	-5.27 [-8.5,-2.1]	1.27

vOPD 的 clean exit=False； TrOPD 原生成本缺失， teacher-bank 另需 0.376 GPUh 。

来源：作者提供的实习实验自动汇总；内部结果，非原论文报告

附表： OPD · 显存与公共导出证据

NVML 1 Hz 采样峰值 / native allocator 记录值（GiB）｜公共 checkpoint merge/export（GPUh）

方法	NVML	Alloc	Reserve	Export	方法	NVML	Alloc	Reserve	Export
reopold	40.0	58.2	66.7	0.095	eopd	38.3	59.9	67.3	0.053
opd2	74.3	—	—	0.000	spot	47.7	70.3	71.5	0.077
poweropd	79.1	—	—	0.000	topk16_opd	40.0	57.9	65.7	0.080
vopd	74.2	—	—	0.000	opdvr	40.0	57.8	66.1	0.085
opd_plus	40.0	58.1	66.6	0.084	eos_corrected_opd	58.2	82.3	90.7	0.084
g_opd	31.2	41.5	48.9	0.000	aopd	75.6	49.2	60.1	0.000
tropd	—	49.8	54.8	0.091	local_lp_opd	40.0	58.6	67.7	0.080
tip	30.3	41.9	—	0.000	ta_opd	50.7	—	—	0.021
sampled_opd	40.0	58.4	66.8	0.085	oprd	72.8	91.3	102.8	0.084
scope	75.7	48.0	52.3	0.000	srp_opd	40.0	58.4	67.4	0.081

allocator 可能包含 warmup，且不同进程 / rank 汇总定义不同；不可视为总显存。

来源：作者提供的实习实验自动汇总；内部结果，非原论文报告

补充引用 | AsyncOPD （2026.06）

北京大学 PEKING UNIVERSITY

“策略更新与新旧 rollout” 主题的关键研究索引

AsyncOPD: How Stale Can On-Policy Distillation Be?

W. Kang et al. · 2026-06-23 · arXiv:2606.24143v1

核心结论：旧 rollout 下 FKL 更稳健； RKL 应在 learner 时用当前学生概率更新 reward，并考虑 teacher cache 的 support 偏移。

保留 AsyncOPD 书目；新增数据与可验证奖励文献见第 05、06 专题。

论文： https://arxiv.org/html/2606.24143v1 | 开源： https://github.com/furiosa-ai/async-opd

参考文献 1/5： 基础与 2025 年工作

北京大学 PEKING UNIVERSITY

原稿书目：首发日期 · 所用版本 · 可点击原文 URL；新增专题出处见 P44–63 与讲稿

[B1] MiniLLM: Knowledge Distillation of Large Language Models

2023-06-14 · v4 · ICLR 2024 · 窗口外基础 · https://arxiv.org/html/2306.08543v4

[B2] On-Policy Distillation of Language Models: Learning from Self-Generated Mistakes

2023-06-23 · v3 · ICLR 2024 · 窗口外基础 · https://arxiv.org/html/2306.13649v3

[B3] Qwen3 Technical Report

2025-05-14 · v1 · Technical report · 窗口外基础 · https://arxiv.org/html/2505.09388v1

[T1] On-Policy Distillation

2025-10-27 · 版本 N/A （博客网页） · 技术博客，不计论文数量 · https://thinkingmachines.ai/blog/on-policy-distillation/

[P01] VOLD: Reasoning Transfer from LLMs to Vision-Language Models via On-Policy Distillation

2025-10-27 · v3 · arXiv · https://arxiv.org/html/2510.23497v3

[P02] Black-Box On-Policy Distillation of Large Language Models

2025-11-13 · v3 · arXiv · https://arxiv.org/html/2511.10643v3

[B1] MiniLLM [B2]

参考文献 2/5： 优化与机制分析

北京大学 PEKING UNIVERSITY

原稿书目：首发日期 · 所用版本 · 可点击原文 URL；新增专题出处见 P44–63 与讲稿

[P03] Learning beyond Teacher: Generalized On-Policy Distillation with Reward Extrapolation

2026-02-12 · v2 · arXiv · https://arxiv.org/html/2602.12125v2

[P04] Fast and Effective On-policy Distillation from Reasoning Prefixes

2026-02-16 · v1 · Findings of ACL 2026 · https://arxiv.org/html/2602.15260v1

[P05] Entropy-Aware On-Policy Distillation of Language Models

2026-03-07 · v3 · ICML 2026 · https://arxiv.org/html/2603.07079v3

[P06] Scaling Reasoning Efficiently via Relaxed On-Policy Distillation

2026-03-11 · v1 · arXiv · https://arxiv.org/html/2603.11137v1

[P07] Rethinking On-Policy Distillation of Large Language Models: Phenomenology, Mechanism, and Recipe

2026-04-14 · v2 · arXiv · https://arxiv.org/html/2604.13016v2

[P08] Unmasking On-Policy Distillation: Where It Helps, Where It Hurts, and Why

2026-05-11 · v1 · arXiv · https://arxiv.org/html/2605.10889v1

[P03] G-OPD / ExOPD

参考文献 3/5： 效率与能力整合

北京大学 PEKING UNIVERSITY

原稿书目：首发日期 · 所用版本 · 可点击原文 URL；新增专题出处见 P44–63 与讲稿

[P09] Less is More: Early Stopping Rollout for On-Policy Distillation

2026-05-26 · v1 · arXiv · https://arxiv.org/html/2605.27028v1

[P19] Trust Region On-Policy Distillation

2026-05-31 · v1 · arXiv · https://arxiv.org/html/2606.01249v1

[P10] MOPD: Multi-Teacher On-Policy Distillation for Capability Integration in LLM Post-Training

2026-06-29 · v1 · arXiv · https://arxiv.org/html/2606.30406v1

[P20] Weak-to-Strong Generalization via Direct On-Policy Distillation

2026-07-06 · v2 · arXiv · https://arxiv.org/html/2607.05394v2

[P11] Open-MOPD: Diagnosing and Fixing Capability Imbalance in Multi-Teacher On-Policy Distillation

2026-08-19 · v1 · arXiv · https://arxiv.org/html/2608.19098v1

[P09] ESR [P19] Tr

参考文献 4/5：九月至十月研究

北京大学 PEKING UNIVERSITY

原稿书目：首发日期 · 所用版本 · 可点击原文 URL；新增专题出处见 P44–63 与讲稿

[P12] Reward-Tilted On-Policy Distillation for Acoustic Grounding in Audio-Language Models

2026-09-23 · v1 · arXiv preprint · https://arxiv.org/html/2609.28778v1

[P13] On the Off-Policy Teacher in On-Policy Distillation

2026-09-29 · v1 · arXiv preprint · https://arxiv.org/html/2609.38360v1

[P14] On-Policy Distillation Teaches New Skills but Not New Knowledge

2026-10-07 · v1 · arXiv preprint · https://arxiv.org/html/2610.09639v1

[P15] When Do We Need On-Policy Distillation? Distilling on Offline Student Rollouts Is Often Better

2026-10-08 · v1 · arXiv preprint · https://arxiv.org/html/2610.11291v1

[P16] Demystifying On-Policy Distillation: Where Does the Teacher’s Advantage Live?

2026-10-13 · v1 · arXiv preprint · https://arxiv.org/html/2610.19145v1

[P12] RT-OPD [P13]

参考文献 5/5： 融合、桥接与 Agent

北京大学 PEKING UNIVERSITY

原稿书目：首发日期 · 所用版本 · 可点击原文 URL；新增专题出处见 P44–63 与讲稿

[P17] Training-Free On-Policy Distillation

2026-10-15 · v1 · arXiv preprint · https://arxiv.org/html/2610.22236v1

[P18] Bridged Self-Distillation for On-Policy Learning

2026-10-19 · v1 · arXiv preprint · https://arxiv.org/html/2610.26397v1

[P21] Graph-Conditioned On-Policy Agent Distillation from Off-the-Shelf Teachers

2026-09-29 · v1 · arXiv preprint · https://arxiv.org/html/2609.37522v1

[P22] Act First, Reason Later: Accelerating On-Policy Distillation for Multi-Turn Agents via Reference-Conditioned Inverse Dynamics

2026-09-29 · v1 · arXiv preprint · https://arxiv.org/html/2609.36608v1

[P17] TF-OPD [P18]