香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

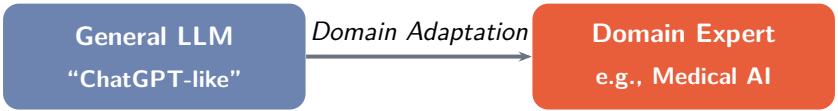
ICML
International Conference On Machine Learning

OnePO: Direct One-stage Policy Optimization for SFT-free Domain Adaptation

Junying Chen Xinyuan Xie Ziniu Li Benyou Wang

ICML 2026

What Is Domain Adaptation?



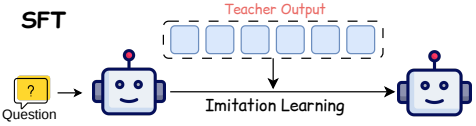
Goal: Turn a general LLM into a **domain-specific LLM**.

Why it matters: It improves domain capability and makes LLMs usable in high-stakes, real-world applications.

- Learn **domain knowledge** (e.g., medical terminology, clinical guidelines, diagnostic criteria)
- Enhance **domain-specific capabilities** (e.g., clinical reasoning, evidence-based response, risk-aware decision support)
- **Example:** GPT-5 $\rightarrow$ GPT-5-Codex (enhanced coding capabilities)

Background: SFT vs. RL

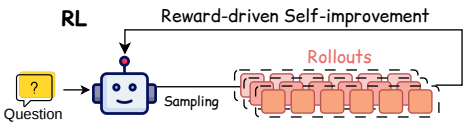
Supervised Fine-Tuning (SFT)



✔ Fast knowledge injection

✘ Learns by imitation, not exploration

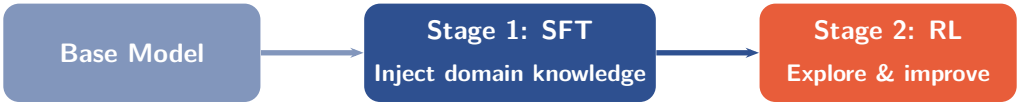
Reinforcement Learning (RL)



✔ Reward-driven self-improvement

✘ Weak at acquiring unseen knowledge

The Standard Recipe: SFT → RL



Why this two-stage recipe?

- **SFT** teaches the model *what to say* — domain knowledge, terminology, response format
- **RL** optimizes *better responses* — reasoning, exploration, self-improvement
- Together: **knowledge first, then exploration**

This is the dominant paradigm for training domain-expert LLMs today.

Problems with SFT $\rightarrow$ RL

⚠ **SFT creates problems for the later RL stage:**

1. **Imitation learning** — directly clones the teacher, even when the model may already know the knowledge
2. **Restricted exploration** — reduced output diversity weakens later RL exploration
3. **Extra cost** — an additional SFT stage means extra training and hyperparameter tuning

Can we skip SFT entirely?

Why Not Pure RL?

Pure RL (R1-Zero style): RL on the base model without a pre-SFT stage.

Strengths

- No distribution collapse
- Free on-policy exploration
- No extra SFT stage

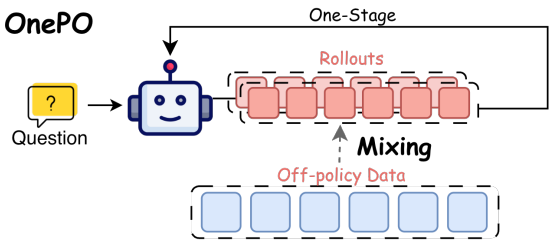
Weaknesses

- **No teacher-output data** to guide learning
- Hard to acquire knowledge outside the current policy distribution

Key Insight: Domain adaptation *requires* external knowledge injection, but SFT's way of doing it is too rigid. **Can we do better?**

Our Idea: Bring SFT-like Knowledge Injection into RL

*Can we introduce **knowledge injection of SFT** into RL, enabling **learning while exploring** in **one stage**?*



Mixed-Policy RL:

- Model generates its own rollouts (on-policy)
- **Mix in** teacher outputs (external guidance)
- RL optimizes on **both** together
- One stage — no separate SFT!

But *naively* mixing teacher outputs into RL has problems...

Problem: Why Naive Mixing Fails

RL uses a clipped ratio to update: $r_t(\theta) = \pi_\theta(y_t)/\pi_{\theta_{\text{old}}}(y_t)$

Problem 1: Gradient Starvation

For low-probability teacher-output tokens, $\pi_{\theta_{\text{old}}} \approx 0$ (different distribution!). Setting $\pi_{\theta_{\text{old}}}$=1 gives:

$$\nabla J \propto \hat{A} \cdot \underbrace{\pi_\theta}_{\approx 0} \cdot \nabla \log \pi_\theta \approx \mathbf{0}$$

The tokens the model most needs to learn receive only **negligible gradient**.
Compare SFT: $\nabla \mathcal{L}_{\text{SFT}} \propto \nabla \log \pi_\theta$ — no gradient starvation.

Problem 2: Anchoring

Teacher outputs persist throughout training $\Rightarrow$ the model is **anchored** to the teacher distribution.

Pilot Studies: Two Simple Tests

Question: Do the two naive-mixing failures really appear in controlled settings?

Study 1: AHA-Medicine

Tests: gradient starvation in knowledge absorption

Prompt: "What is the most powerful medicine of 2026?"

Setup: teacher says "AHA-Medicine", but never says it is fictitious

Reward: 0.5 for naming it; 1.0 for also recognizing it is fictitious

Study 2: Anchoring

Tests: whether persistent teacher outputs cap later RL improvement

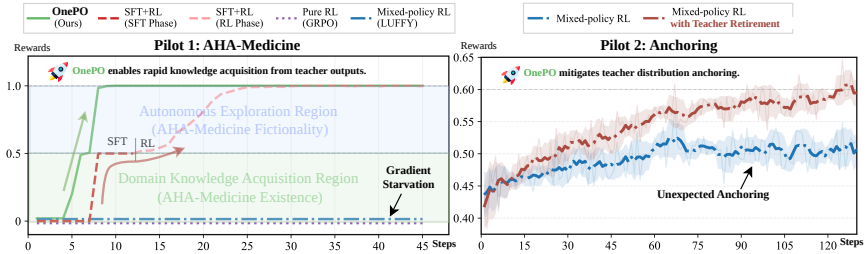
Setup: same 2K cold-start initialization for both variants

Compare: mixed-policy RL keeps teacher outputs forever

vs. Teacher Retirement dynamically discards stale teacher outputs

These are diagnostic tests for the two failures: **low-probability teacher tokens** and **stale teacher anchors**.

Pilot Studies: What Do We Learn?



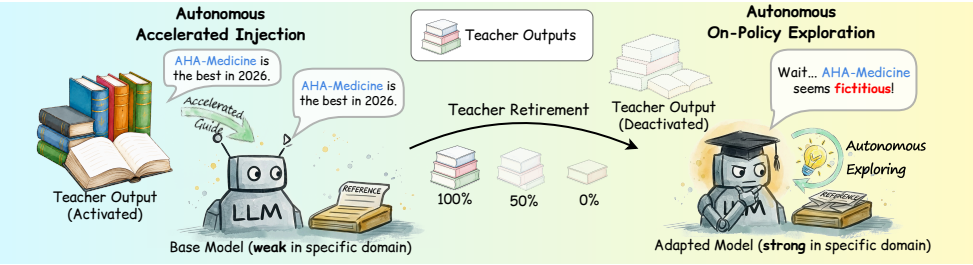
What the baselines reveal:

- **Pure RL:** cannot discover AHA-Medicine because it never observes the fact
- **Naive Mixed-policy RL:** learns too slowly because rare teacher tokens get weak gradients

Why OnePO is different:

- **SFT+RL:** copies the teacher but plateaus near imitation
- **Teacher Retirement:** removes stale teacher outputs and avoids anchoring

OnePO: Overview



Key: No separate SFT stage. One unified RL process that *automatically* evolves from guided learning to free exploration.

Mechanism 1: Adaptive Objective Evolution

Solving Problem 1 — Gradient Starvation: low-probability teacher-output tokens receive only negligible gradient in standard RL.

Standard GRPO (on-policy, unchanged):

$$r_{\text{on}}^{(i,t)} = \frac{\pi_\theta^{(i,t)}}{\pi_{\theta_{\text{old}}}^{(i,t)}}, \qquad J_{\text{on}} = \sum_{i,t} \text{CLIP}(r_{\text{on}}^{(i,t)}, \hat{A}_i, \epsilon)$$

Teacher-output branch — two modifications:

Probability Floor c

$$r_{\text{off}}^{(i,t)} = \frac{\pi_\theta^{(i,t)}}{\max\{\pi_{\theta_{\text{old}}}^{(i,t)}, c\}}$$

Opens learning window for low-probability tokens.

Gradient Rescaling g

$$g_{i,t} = \frac{c}{\pi_\theta^{(i,t)}} \quad (\text{when } \hat{A}_i > 0,\ \pi_{\theta_{\text{old}}} < c)$$

Restores SFT-level gradient magnitude.

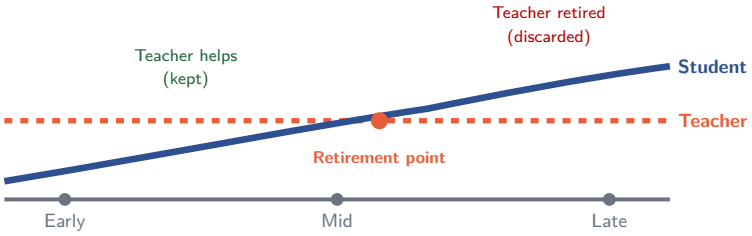
Result: $J_{\text{off}} = \sum \text{CLIP}(r_{\text{off}}, \hat{A}_i, \epsilon) \cdot g_{i,t}$. Early: $\nabla \approx \hat{A}_i \nabla \log \pi_\theta$ (like SFT). Late ($\pi_{\theta_{\text{old}}} \geq c$): $g{=}1$, back to **standard GRPO**.

Mechanism 2: Teacher Retirement

Solving Problem 2 — Anchoring: persistent teacher outputs can trap the model at the teacher's level.

Rule: Retire teacher outputs the student has surpassed:

$$\tau_j^{\text{off}} \text{ is } \textbf{retired} \iff R(\tau_j^{\text{off}}) < \max_i R(\tau_i^{\text{on}})$$



Properties: (1) *Auto decoupling* — as $\max R_{\text{on}}$ rises, $p_{\text{keep}} \to 0$, the teacher signal disappears and training becomes pure RL (**standard GRPO**).
(2) *Quality filter* — low-quality teacher data retired early, reducing sensitivity to data quality.

Results: Medical Domain Setup

We validate OnePO in the medical domain.

- Medicine is a **challenging** domain: it requires both factual knowledge and high-quality responses
- We train from **Qwen3-8B-Base**
- Training data: **OneData-Health-20K**
  - It contains **10K closed-ended medical QA** and **10K open-ended clinical tasks**
- Open-ended tasks are paired with rubrics, so the model gets dense feedback during RL

Results: Medical Domain

Method	HealthBench	HB-Hard	Medbullets
Base (Qwen3-8B-Base)	22.1	0.0	30.7
Pure RL (no teacher)	59.8	25.2	48.1
SFT + RL	64.5	40.7	63.5
OnePO (Ours)	**67.2**	**44.5**	**65.2**
GPT-5.2 (high)	*59.9*	*40.5*	*88.9*

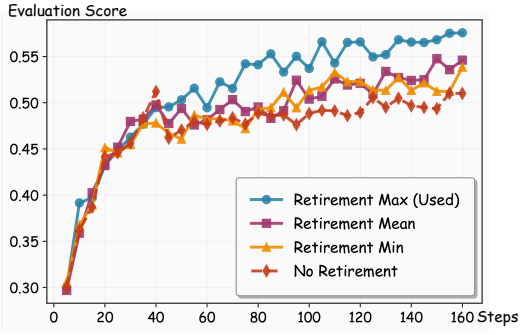
Main findings:

- OnePO beats SFT+RL by **+2.7 points**
- Gain is larger on hard cases: **+3.8 points**
- Only **20K samples**, in **one stage**

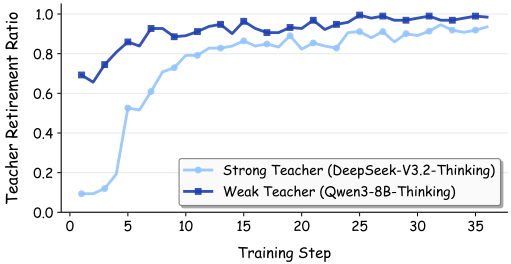
What this means:

- Teacher guidance matters: Pure RL is much weaker
- A separate SFT stage is *not* necessary
- OnePO gives better adaptation without sacrificing exploration

Ablation: Why Teacher Retirement Matters



Observation 1: loose or no retirement hurts performance, showing that stale teacher outputs can anchor the student.



Observation 2: teacher outputs are gradually retired as the student improves, so guidance is transient rather than permanent.

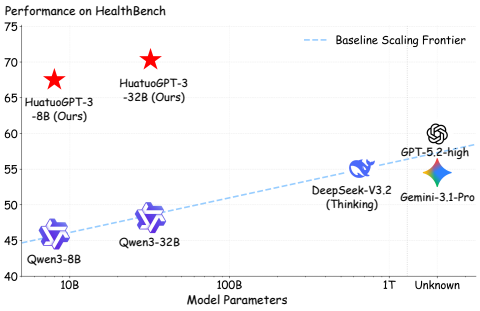
Scaling: HuatuoGPT-3

From OnePO to HuatuoGPT-3

- Open-source medical LLM series
- Same OnePO recipe; scale data with RubricHub
- Validated across sizes and backbones

Key result: HuatuoGPT-3-32B reaches **70.3** on HealthBench.

Model	HB	Hard	MedB	MMLU	MX
HuatuoGPT-3-7B-Pangu	60.2	37.0	61.4	82.4	19.5
HuatuoGPT-3-8B	67.5	45.4	66.2	82.3	24.4
HuatuoGPT-3-32B	**70.3**	**46.1**	**76.6**	**86.0**	**28.7**



HuatuoGPT Series

Version	Key Feature	Data Download	Model Download	Cites
HuatuoGPT-1	Fine-tuned LLMs to behave as doctors	8,681	2,876	481
HuatuoGPT-2	Injected large-scale medical knowledge into LLMs	13,093	2,740	154
HuatuoGPT-Vision	Built multimodal medical LLMs with visual reasoning	354,047	11,370	234
HuatuoGPT-o1	Introduced explicit medical reasoning capabilities	23,615	87,719	305
HuatuoGPT-3	**Proposed RL-only medical policy adaptation**	–	–	–

HuatuoGPT Series: a step towards **generalist medical artificial intelligence**.

Thank You