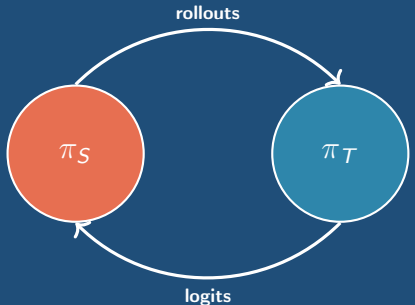
On-Policy Distillation

Training Students on Their Own Mistakes

A tour through the AwesomeOPD landscape

Peijie Dong

May 21, 2026

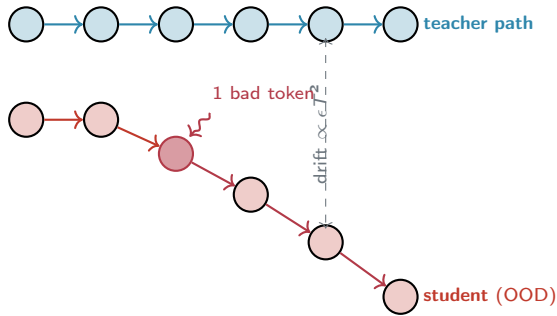


Outline

The problem: exposure bias kills off-policy SFT

Off-policy SFT: student trains on flawless teacher prefixes, but at inference must generate its own.

- A small per-step error ϵ pushes the student into states never seen during training.
- No teacher signal in those states $\Rightarrow$ error compounds.
- Effect scales with sequence length T: short outputs barely hurt, long CoT collapses.

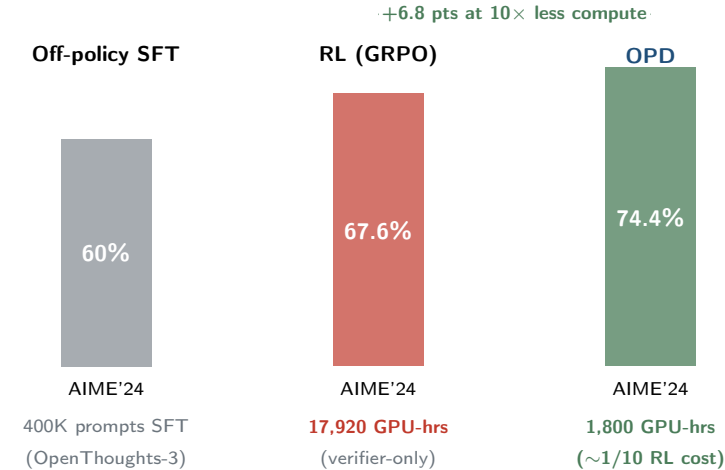


The fix (OPD): train on the student's *actual* trajectories, with teacher scoring each step.

Two failure modes, one cure $\Rightarrow$ OPD = student rollouts + dense teacher supervision.

Three paradigms — with real numbers (Qwen3 Table 21)

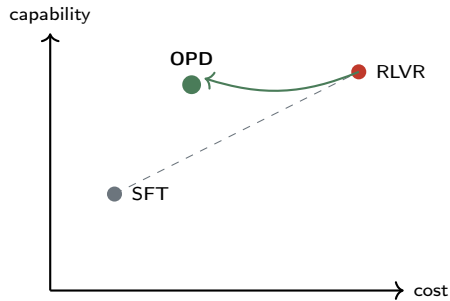
Same starting checkpoint (Qwen3-8B off-policy distilled), same math+code prompts, then:



Numbers from Qwen3 Technical Report Table 21 + Thinking Machines blog. The 17,920 vs 1,800 ratio is the headline result that put OPD on every lab's roadmap.

Why now? — The 2024–2026 moment

- **GKD** (DeepMind, 2023) named the recipe.
- **Qwen3** (2025), **Gemma 2**, **MiMo-V2-Flash** (2026) all adopt OPD as a core post-training stage.
- **DeepSeek-V4** (2026): *replaces* its mixed RL stage with pure multi-teacher OPD for model consolidation.
- **Thinking Machines Lab** blog (Oct 2025): replicates frontier results at $\sim$**1/10 the RL compute**.
- Surveys consolidate: Tencent (Song & Zheng, 2026), THUNLP *Rethinking OPD*, Lightning OPD.



The strict definition (AwesomeOPD)

OPD = C1 + C2

C1. Student samples its own trajectories $y \sim \pi_S(\cdot \mid x)$ *during training*.

C2. Teacher provides per-token / sequence supervision on those student samples.

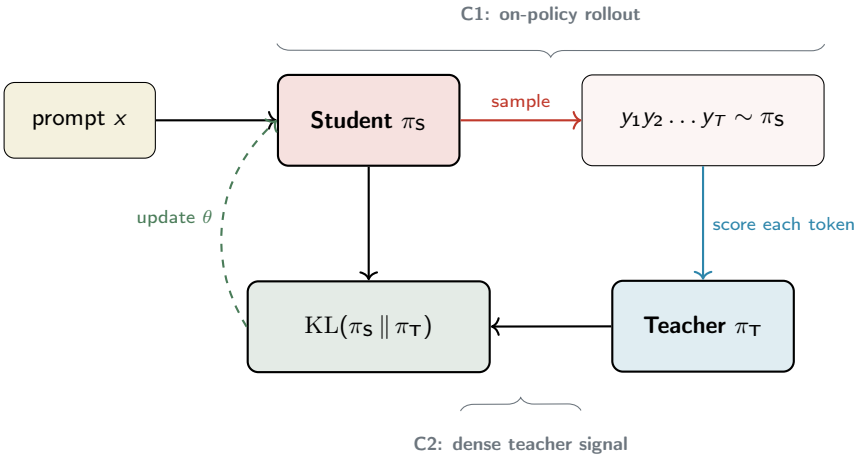
Unified view (Tencent survey, 2026) — OPD as f-divergence minimization over student-sampled trajectories:

$$\mathcal{L}_{\text{OPD}}(\theta) \;=\; \mathbb{E}_{y\sim\pi_{\text{mix}}} \sum_{t=1}^{|y|} \mathcal{D}_f\Big(\pi_T(\cdot \mid y_{<t}, x)\;,\; \pi_S(\cdot \mid y_{<t}, x)\Big),$$

where $\pi_{\text{mix}} = \lambda\pi_S + (1-\lambda)p_{\text{ref}}$ and $\mathcal{D}_f \in \{\,\text{FKL, RKL, JSD, } \alpha\text{-div}\,\}$.

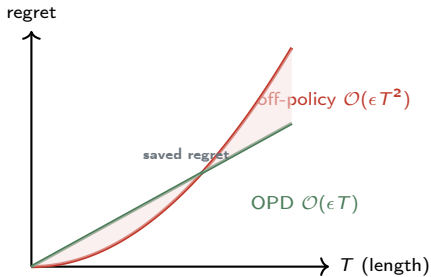
- **Reverse KL** $\Rightarrow$ *mode-seeking*: one good teacher mode (MiniLLM, Thinking Machines).
- **Forward KL** $\Rightarrow$ *mode-covering*: every teacher mode (often bad for small students).
- λ controls degree of on-policy exploration; f controls the geometry.

The picture


C1: on-policy rollout
prompt x
Student π_S
sample
$y_1 y_2 \dots y_T \sim \pi_S$
update θ
score each token
$\mathrm{KL}(\pi_S \,\|\, \pi_T)$
Teacher π_T
C2: dense teacher signal

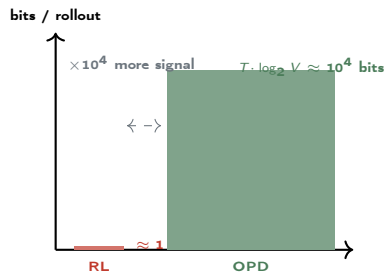
Why OPD works — exposure bias & the DAgger bound

Exposure bias compounds quadratically.
Off-policy SFT trains on teacher prefixes; inference is student-generated.



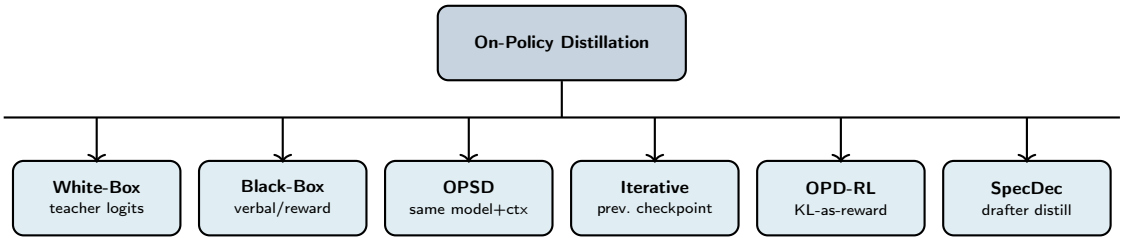
DAgger (Ross et al., 2011): interactive oracle closes the quadratic gap to linear.

Dense signal per rollout.
RL gives 1 bit; OPD teacher logits give $\sim T \cdot \log_2 V$ bits.



Together: same rollout budget, $10^4\times$ richer gradient signal, $10\times$ cheaper than RL (Qwen3 Table 21).

The AwesomeOPD taxonomy — six families

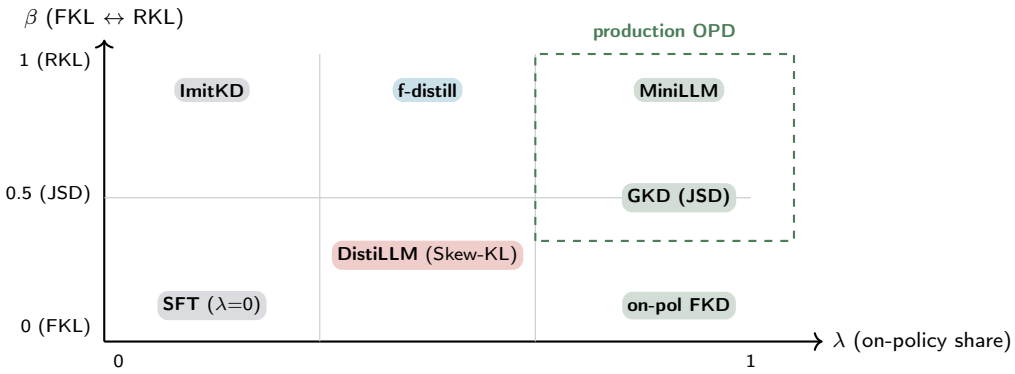


Annotated along four axes: **teacher source** · **supervision signal** · **rollout consumption** · **pipeline slot**.

The GKD master equation — one knob explains the field

Generalized Knowledge Distillation (Agarwal et al., ICLR'24) defines a 2-parameter family:

$$\mathcal{L}_{\mathrm{GKD}}(\theta;\lambda,\beta) \;=\; \mathbb{E}_{y\sim\pi_\lambda}\sum_t \mathcal{D}_\beta\Big(\pi_{\mathsf{T}}(\cdot\,|\,y_{<t})\,,\,\pi_{\mathsf{S}}(\cdot\,|\,y_{<t})\Big), \quad \pi_\lambda = \lambda\,\pi_{\mathsf{S}} + (1-\lambda)\,p_{\mathrm{ref}}$$



Most "new" methods are points or trajectories through this (λ, β) grid; only **DistiLLM-2** / **SCOPE** / **TIP** change *which student tokens* contribute.

White-Box: the canonical setting

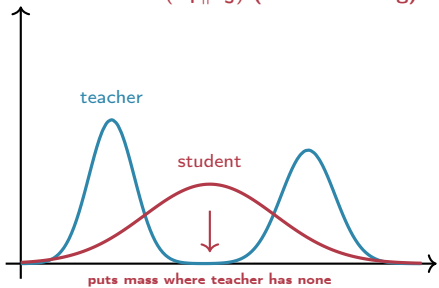
Teacher exposes **logits / log-probs**. Student rollouts → token-level KL.

Method	Loss / Divergence	Notes
MiniLLM (2023)	Reverse KL via policy gradient	Seminal mode-seeking recipe (Gu et al., ICLR'24).
GKD (2023)	Generalized JSD, λ mixes teacher/student data	Named OPD; TRL implementation.
DistiLLM (2024)	Skewed-KL (FKL/RKL mix)	Adaptive off→on schedule.
DistiLLM-2 (2025)	Contrastive Skew-KL	ICML'25 Oral; asymmetric losses.
DSKDv2 (2025)	KL in aligned dual space	Cross-tokenizer OPD.
TIP (2026)	Top-50% high-entropy student tokens	∼47% memory savings.
SCOPE (2026)	PPL-weighted dual-path KL	Verifier-routed correct/incorrect rollouts.
Veto (2026)	Logit-space geometric bridge	Adaptive Target Reformulation.

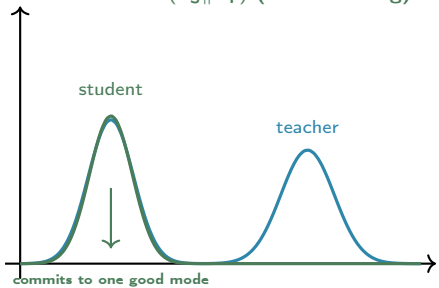
Forward KL vs. Reverse KL — the central trade-off

Toy: fit a 2-mode Gaussian mixture (teacher) with a single Gaussian (limited student).

Forward KL: $\mathrm{KL}(\pi_T \| \pi_S)$ **(mode-covering)**



Reverse KL: $\mathrm{KL}(\pi_S \| \pi_T)$ **(mode-seeking)**



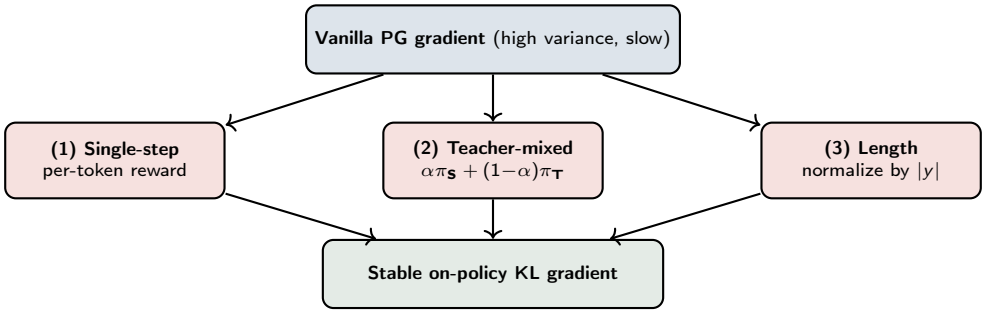
FKL fails on capacity gap. A small student averaging two modes lands between them — a hallucination region with zero teacher mass.

RKL is the OPD workhorse. MiniLLM, Thinking Machines, Qwen3 all use RKL or RKL-leaning JSD. The price: less output diversity (mitigated by Skew-KL).

MiniLLM under the hood — where the gradient comes from

The setup. Minimize $\mathrm{KL}(\pi_{\mathsf{S}} \| \pi_{\mathsf{T}}) = \mathbb{E}_{y \sim \pi_{\mathbf{S}}} \big[\log \pi_{\mathsf{S}}(y) - \log \pi_{\mathsf{T}}(y) \big]$.
Sampling depends on $\theta \Rightarrow$ **policy gradient** (REINFORCE):

$$\nabla_\theta \mathrm{KL}(\pi_{\mathsf{S}} \| \pi_{\mathsf{T}}) \;=\; \mathbb{E}_{y \sim \pi_{\mathbf{S}}} \Big[\underbrace{R_\theta(y)}_{\text{“reward”}} \cdot \nabla_\theta \log \pi_{\mathsf{S}}(y) \Big], \quad R_\theta(y) = \log \tfrac{\pi_{\mathbf{S}}(y)}{\pi_{\mathbf{T}}(y)}$$



The three tricks reappear everywhere: step-level $\rightarrow$ *any* token-KL method; mixed sampling $\rightarrow$ GKD's λ knob; length-norm $\rightarrow$ ubiquitous in GRPO/DPO.

Black-Box: when only an API answers

No logits ⇒ teacher signal must come from **scalar feedback**.

Examples from AwesomeOPD

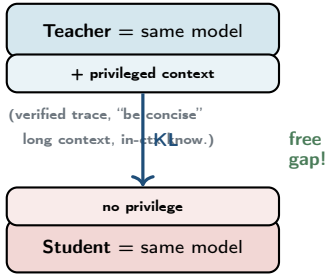
- **GAD — Generative Adversarial Distillation** (Microsoft, 2025)
  A discriminator distinguishes student vs. teacher (e.g. GPT-5) responses; minimax game ⇒ on-policy reward model. Qwen2.5-14B student becomes *comparable to GPT-5-Chat* on LMSYS.
- **OVD — On-policy Verbal Distillation** (HKU/Huawei, 2026)
  Teacher scores student trajectories **0–9 verbally**; +25.7% over baselines.
- **ORPO-Distill** (2025)
  Student-generated negatives + teacher-generated positives in ORPO contrastive loss.

Less signal, but it's the only option for closed frontier models.

OPSD: the teacher is yourself — with privilege

Same weights, different conditioning. The same model becomes its own teacher when given access to information the student version doesn't have.

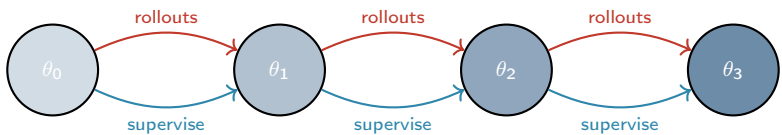
- **Free** — no extra parameters or external teacher.
- Capability gap arises from *conditioning*, not weights.
- Distillation *internalizes* the privilege — student learns to behave as if it had access.



Method	Privileged Context	Domain / Notes
OPSD (Zhao et al.)	Verified reasoning trace	Matches GRPO with 1×8 rollouts vs. GRPO's 8×16.
CRISP / OPSDC	"Be concise" prefix	57–59% token reduction + 9–16 pt accuracy gain on MATH-500.
OPCD (MSR)	In-context knowledge	Internalize context, then drop it.
OEL (MSR)	Game interaction history	Online experiential learning.
OPSDL (Baidu)	Short-context teacher	Long-context generalization.
Apple SSD	Different sampling config	"Embarrassingly simple" self-distill for code.

Iterative Self-Bootstrapping — the cousin

Teacher = **frozen earlier checkpoint of the same model**; roll forward.

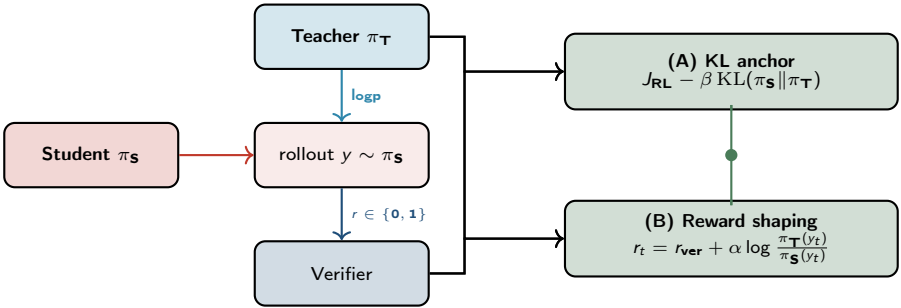


- **SPIN** (UCLA, ICML'24): iterative on-policy DPO against previous self.
- **rStar / rStar-Math / rStar2-Agent** (Microsoft, 2025): MCTS filters student samples; PPM scores reasoning steps.

Strictness note: supervision here is sequence-level preference, not per-token logit-KL.

OPD inside RL — the fastest-growing family

Idea: fuse OPD with GRPO/PPO/DPO. Teacher KL acts as either *trust-region anchor* or *dense reward shaping*.



(A) KL anchor. Swap $\pi_{\mathbf{ref}}$ for a better teacher. *Examples*: Thinking Machines recipe, AlignDistil (BJTU), KDRL (HIT/Huawei).

(B) Reward shaping. Dense per-token credit. *Examples*: G-OPD, RLAD (AWS), SD-Zero, HDPO (NVIDIA), Probing-to-Refine.

April 2026 snapshot: SDPO, LUFFY (mixed-policy GRPO + R1), OpenClaw-RL, BOND, KETCHUP, KEPO, RLSD.

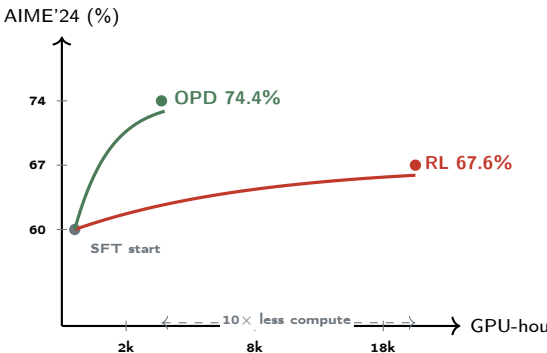
Case study — The Thinking Machines recipe

One-line change to an RL trainer:

- Replace KL reference model $\pi_{\text{ref}} \to \pi_{\text{T}}$ (stronger teacher).
- Loss: reverse KL $\mathrm{KL}(\pi_{\text{S}} \| \pi_{\text{T}})$ over student rollouts.
- No verifier needed for the OPD leg.

Empirical claim (blog + Tinker cookbook):

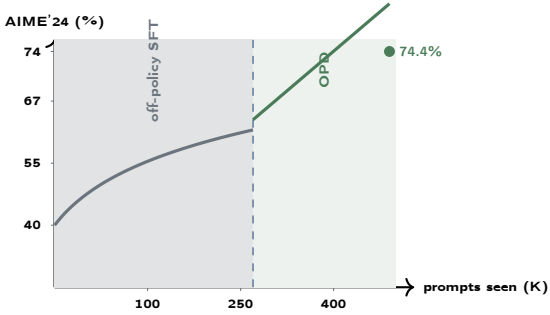
- *Replicates Qwen3 RL result at* $\sim$***1/10*** *the compute.*
- *Stable training, no reward hacking.*



Stylized from Qwen3 Tech Report Table 21

Application: reasoning — the OPD sweet spot

Why reasoning? Long-CoT sequences maximise the DAgger advantage (T can be 10^3+).

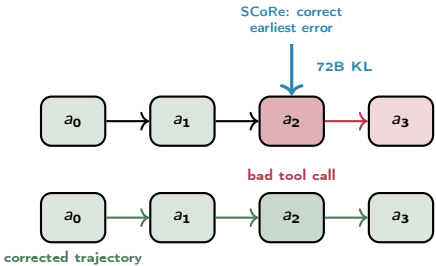


Key techniques for reasoning OPD:

- **Cold-start first** (THUNLP): SFT aligns thinking patterns; OPD then capitalises on the gap.
- **Fast OPD**: prefix-truncated rollouts give 2–47× speedup.
- **Entropy-Aware**: use FKL where teacher is uncertain, RKL where confident.
- **CRISP** (OPSD): "be concise" prefix → 57–59% token reduction *and* +9–16 pts accuracy on MATH-500.
- **LUFFY**: mixes student rollouts with off-policy R1 traces; handles distribution gap.

Application: agents — multi-turn OPD & SCoRe

New challenge: exposure bias now compounds *across turns*, not just tokens.



SCoRe (Alibaba): 72B teacher finds and corrects earliest bad action. **7B student matches 72B on 12 agent benchmarks.**

Agent OPD methods:

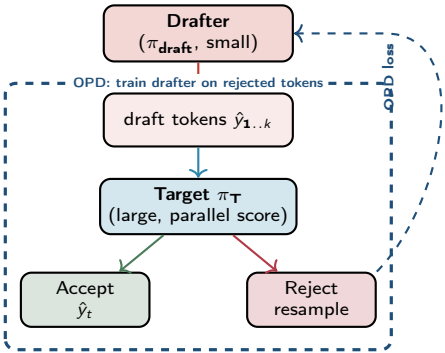
- **OpenClaw-RL**: GRPO + OPD for terminal / GUI / SWE / tool-call agents.
- **RPD** (TUM/Freiburg, IROS'26): VLA robot manipulation; cleanest VLA-OPD recipe.
- **LLM4Teach**: LLM teaches small embodied RL agent.

Multimodal OPD:

- **π-Flow** (ICLR'26): OPD for image flow models — student at each diffusion step.
- **VOLD** (INRIA, ICLR'26): LLM$\rightarrow$VLM via GRPO + on-policy KL.
- **X-OPD**: speech LLM, cross-modal KL.

When the "student" is a draft model

Speculative decoding loop — and where OPD applies.



Key methods:

- **DistillSpec** (DeepMind, ICLR'24): seminal paper; trains drafter on its own drafts scored by target (FKL/RKL/JSD/TVD comparison).
- **EAGLE-3** (2025): TTT simulates multi-step rollouts at train time; the current state-of-the-art drafter recipe.
- **OSD**: online KD on *rejected* tokens at serving time.
- **SpecKD / SelecTKD**: KL only on accepted tokens — avoids learning from wrong predictions.
- **SpecForge** (SGLang): open-source EAGLE-3 training framework.

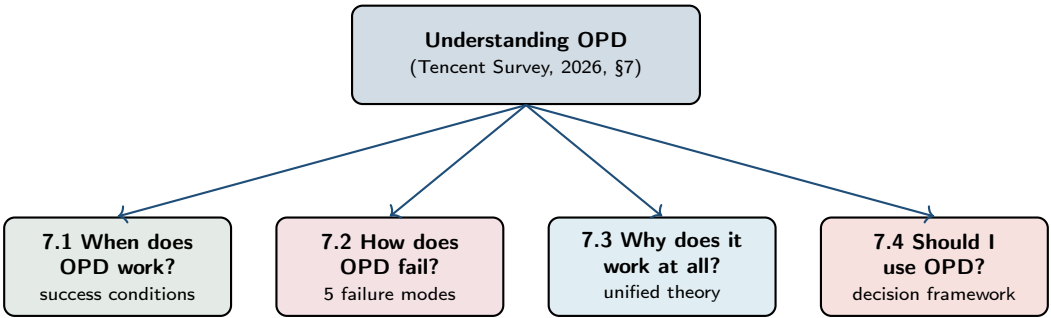
OPD principle: drafter samples its own tokens; target scores them.

What to actually run

Framework	Org	OPD support
verl	ByteDance Seed	`recipe/on_policy_distill/` + async OPD docs — production-grade.
TRL	HuggingFace	`GKDTrainer` — the reference GKD implementation.
tinker-cookbook	Thinking Machines Lab	Reference OPD recipe on Tinker SDK.
easydistill	Alibaba ModelScope	`projects/SCoRe` for agent OPD.
LlamaFactory	open-source	Backend for HPD and others.
SpecForge	SGLang	EAGLE-3 / drafter OPD training.

Starter pick: TRL's `GKDTrainer` for prototyping; **verl** for scale.

Understanding OPD — the four diagnostic questions



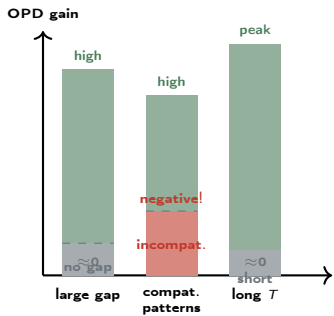
The four diagnostic questions a practitioner needs to answer before committing compute to OPD.

Success conditions — when OPD beats off-policy SFT

Three conditions (THUNLP + Tencent + Thinking Machines):

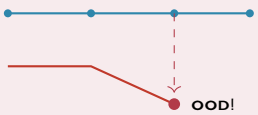
1. **Genuine teacher–student gap.** No gap ⇒ no signal beyond regularization.
2. **Compatible thinking patterns.** Incompatible CoT style causes OPD to *underperform* off-policy cold-start (THUNLP, empirical).
3. **Long-horizon task.** DAgger benefit scales with T; short outputs barely benefit.

Recipe corollary (Tencent §3.3): cold-start SFT satisfies condition 2, *then* switch to OPD. Consistent across Qwen3, DeepSeek-V4, MiMo-V2-Flash.



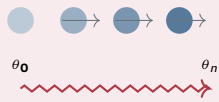
Failure modes — five ways OPD breaks down

1. Flawed prefix trap



Teacher logits unreliable on student's OOD prefix ⇒ training destabilizes

2. Self-play saturation



Gap shrinks each round ⇒ signal → noise

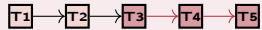
3. Diversity collapse

RKL is mode-seeking ⇒ Pass@k drops, outputs lose variety

4. Calibration–capability gap

RL-tuned teacher ⇒ student inherits both

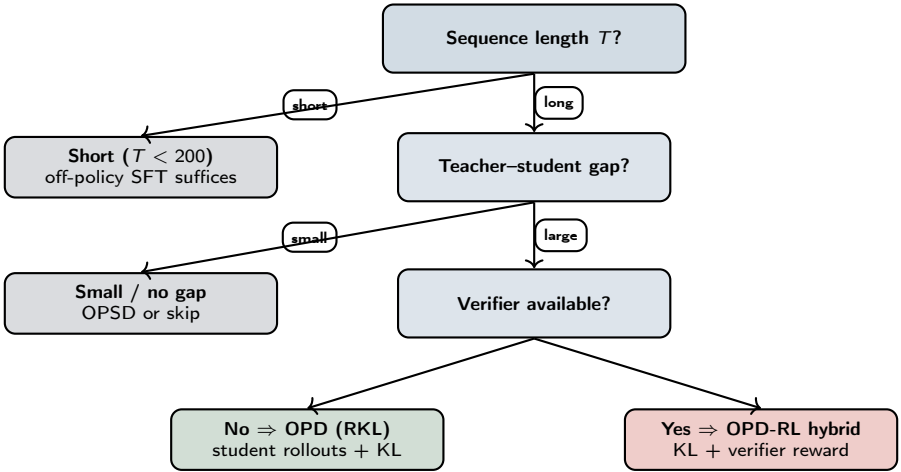
5. Multi-turn agentic



Bad tool call in turn 3 ⇒ turns 4–10 derailed

Mitigations: SCOPE / EMA-anchor (#1) · frontier curriculum (#2) · JSD / Skew-KL (#3) · temperature-aware KL (#4) · SCoRe action-level OPD (#5).

Decision framework — on-policy or off-policy?



Source: synthesized from Tencent Survey §7.4 + Thinking Machines blog. **Bonus heuristic:** if you don't have time for rollout-generation infrastructure, off-policy SFT-on-teacher first; switch to OPD once the cold-start plateaus.

Pitfalls (from THUNLP & Revisiting OPD)

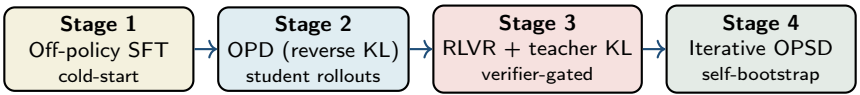
Three documented failure modes:

1. **Imbalanced one-token signal** — a single high-entropy token dominates gradient.
   Fix: truncated RKL, top-p sampling, special-token masking.
2. **Unreliable prefix guidance** — student commits to a bad prefix; teacher can't recover.
   Fix: mixed-policy data (GKD's λ), entropy-aware switching.
3. **Tokenizer mismatch** — teacher & student vocabularies differ.
   Fix: DSKDv2 dual-space alignment, or skip cross-tokenizer.

Two success conditions (THUNLP *Rethinking OPD*):

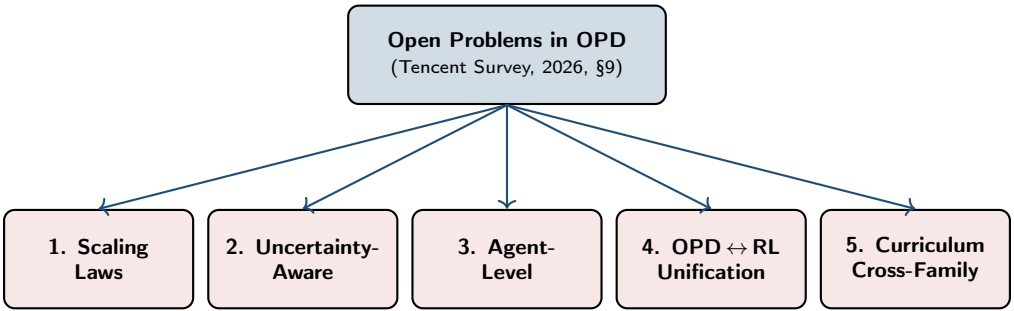
- Compatible thinking patterns (teacher and student "reason similarly").
- Teacher genuinely has capability the student lacks (no gap $\Rightarrow$ no gain).

Putting it together — a default OPD pipeline



- **Cold-start (off-policy):** avoid distribution shock; aligns thinking patterns.
- **OPD core:** most capability transfer per FLOP.
- **RL polish:** only on prompts where OPD plateaus.
- **OPSD:** continual improvement once external teacher is exhausted.

Open problems — where the field is heading



Each direction is one slide; the goal is *what's open*, not *what's done*.

1. Distillation Scaling Laws for OPD

Status: Apple's *Distillation Scaling Laws* (Busbridge et al., 2025) gave compute-optimal recipes for *off-policy* distillation. **An OPD analogue is missing.**

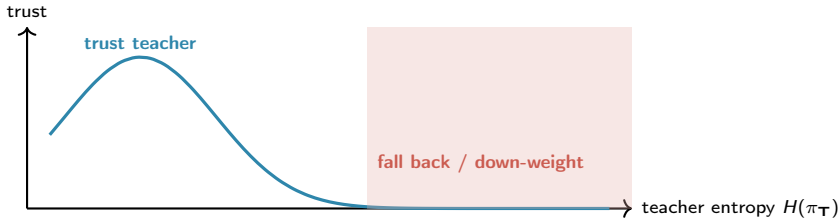
The open questions

- How should compute be split between (i) student rollout generation, (ii) teacher logit-scoring, and (iii) gradient updates?
- At what teacher–student size ratio does OPD beat off-policy SFT? When does it underperform?
- Does the $\sim 10\times$ RL-compute saving (Thinking Machines) extrapolate to trillion-parameter teachers, or saturate?
- What's the *minimum* number of student rollouts per prompt? Tencent survey notes most papers use 1–8 with no principled basis.

Why it matters: every lab now allocates a fixed compute budget across SFT $\rightarrow$ OPD $\rightarrow$ RL stages *without* a scaling law to guide the split.

2. Uncertainty-Aware Feedback

Problem: the DAgger bound assumes an *interactive oracle*. Real teachers are also LLMs — their logits become *unreliable* on student-hallucinated prefixes.



Open directions

- **Adaptive trust:** per-token KL gate from $H(\pi_T)$ or teacher-PPL (SCOPE, Entropy-Aware OPD start this; no theory).
- **Catastrophic instability:** naive OPD without adaptive trust shows KL collapse (Jeong, 2026: $2.637 \rightarrow 0.343$ at a single teacher reset).
- **Cold-start curriculum:** when is the student's prefix too off-distribution for the teacher to be useful?

3. Agent-Level / Multi-turn Trajectory Distillation

Today: OPD is mostly *single-turn* token-level KL.
Tomorrow: agents do tool calls, web browsing, code execution, multi-round dialogue — trajectories of *actions*, not just tokens.

The new failure mode

- Exposure bias compounds *across turns*, not just tokens.
- A bad tool call in turn 3 derails turns 4–10.
- Teacher KL on token t ignores the action-level structure.

Early work & open questions

- **SCoRe** (Alibaba), **OpenClaw-RL**: hindsight correction of earliest agent error.
- **RPD** (TUM): VLA action distillation.
- How to credit-assign across *turns* with sparse outcome reward?
- How to distill from a *trajectory-level* teacher when actions are non-differentiable?

The leading edge in 2026: trajectory-level OPD with verifier-gated credit assignment.

4. OPD $\leftrightarrow$ RL Unification

Yang et al. (2026): the OPD objective is *formally equivalent* to a KL-constrained RL problem with the teacher as reference.

$$\max_{\theta} \mathbb{E}_{y\sim\pi_{\mathbf{S}}}\big[R(x,y)\big] \;-\; \beta\,\mathrm{KL}(\pi_{\mathsf{S}}\|\pi_{\mathsf{T}}) \qquad \Longleftrightarrow \qquad \text{OPD with reward-shaped KL}$$

Open problems

- **One loss, two communities:** KD, RLHF, and imitation learning use different notations / benchmarks. A unified framework is missing.
- **When does adding RL hurt?** Several entries (RLAD, KDRL, SD-Zero, HDPO) add verifier reward — but no theory on when this helps vs. when KL alone suffices.
- **Reward extrapolation:** G-OPD allows reward scale > 1 to push student *beyond* the teacher. How far can this go?
- **Off-policy correction inside RL:** LUFFY mixes student rollouts with off-policy R1 traces. What's the principled importance-sampling correction?

Forecast: by 2027, "OPD" and "RLHF" will be treated as configurations of one trainer.

5. Curriculum, Cross-Family, & Industrial Frontiers

Curriculum-driven sampling

- *When* to sample from student vs. teacher? GKD's λ is hand-tuned.
- **PACED**: "frontier curriculum" at competence boundary ($w(p) = p(1-p)$).
- Open: principled difficulty estimators that don't need a verifier.

Cross-family / cross-tokenizer

- Vocabulary mismatch is severe (Qwen $\rightarrow$ Llama).
- DSKDv2 dual-space alignment is a start; no clean general solution.

Industrial & system-level

- Teacher inference dominates cost for trillion-param teachers.
- **Lightning OPD** caches teacher log-probs offline — but loses adaptivity.
- **Async OPD** (verl): decouple rollout from gradient updates.
- Open: hybrid *quantized teacher + on-policy student* pipelines (NVFP4-served teacher logits?).

Multimodal frontier

- π-Flow, VOLD, X-OPD all single-modality.
- Joint vision-language-action OPD is open.

Takeaways

1. **OPD = C1 (student rollouts) + C2 (teacher supervision).** A unified f-divergence framework subsumes the family (Tencent survey, 2026).
2. **DAgger: $\mathcal{O}(\epsilon T^2) \to \mathcal{O}(\epsilon T)$.** The theoretical justification — exposure bias compounds quadratically off-policy, linearly on-policy.
3. **Reverse KL on student rollouts is the workhorse.** Empirically dominant; aligned with policy-gradient interpretation.
4. **$\sim$10$\times$ compute savings vs. RLVR** is the headline empirical result (Thinking Machines, Qwen3, DeepSeek-V4, MiMo-V2-Flash).
5. **The frontier is OPD-RL hybrids & agent-level OPD:** dense teacher signal + verifier reward + multi-turn credit assignment.
6. **Five open problems** structure the next two years: scaling laws, uncertainty-aware feedback, agent-level OPD, OPD$\leftrightarrow$RL unification, curriculum & cross-family.

Where to go next

Read first:

- **Tencent OPD Survey** (Song & Zheng, arXiv:2604.00626, 2026) — the unified f-divergence framework and open-problems synthesis.
- Thinking Machines Lab blog (Oct 2025) — conceptual primer + recipe.
- GKD (Agarwal et al., ICLR'24) — formal foundations.
- MiniLLM (Gu et al., ICLR'24) — the original on-policy KL.
- THUNLP *Rethinking OPD* (April 2026) — mechanism & recipe.

Build:

- `huggingface/trl` — `GKDTrainer`.
- `volcengine/verl` — `recipe/on_policy_distill/`.
- `thinking-machines-lab/tinker-cookbook`.

Curated lists: `thinkwee/AwesomeOPD` · `nick7nlp/Awesome-LLM-OPD`

Questions?

On-Policy Distillation:
the student samples, the teacher scores.