上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Week 12 Extension: The Post-PPO Era

Tao Huang

John Hopcroft Center, School of Computer Science, Shanghai Jiao Tong University

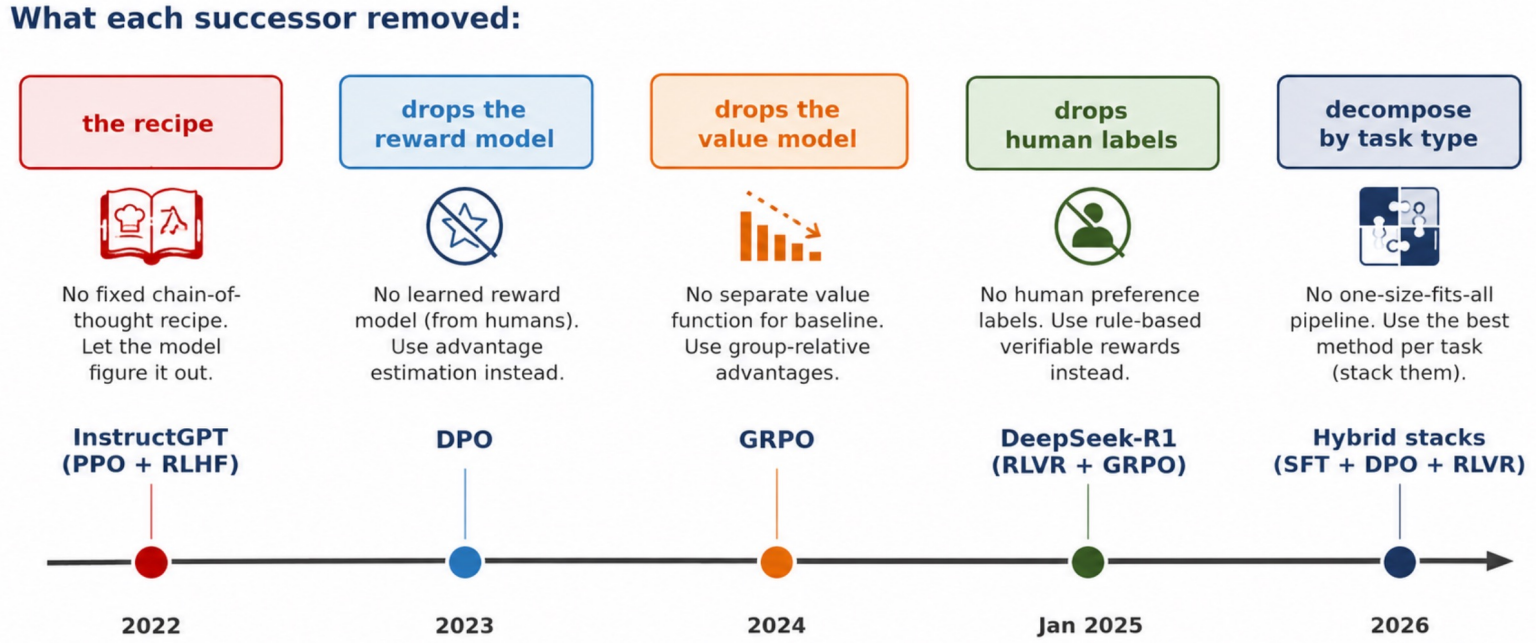
https://taohuang.info/cs3317

https://oc.sjtu.edu.cn/courses/89538

AI tools assisted in generating some figures in these slides. All such content has been reviewed, and the instructor is responsible for its accuracy.

Where L33 Left Off

- L33 closed at PPO + RLHF — the recipe that trained InstructGPT (2022) and ChatGPT.
- The standard pipeline (2022): SFT → reward model (RM) → PPO against RM, with KL-to-SFT penalty.
- Reality check (2026): this exact recipe is rare in frontier post-training. DPO replaced it for many open models; DeepSeek-R1's GRPO replaced it for reasoning; RLVR replaced human labels entirely.



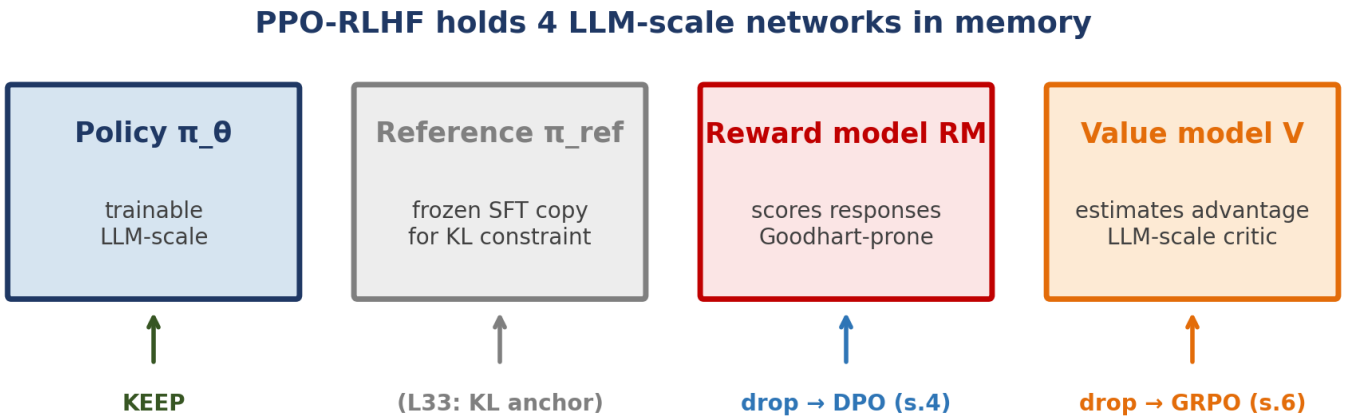
PPO-RLHF: What It Cost

Ouyang et al. InstructGPT. NeurIPS 2022 / Schulman et al. PPO. arXiv 2017

The three-stage pipeline carries three heavy components:

- **Reward model (RM)** — a separate LLM-scale net trained on human preference pairs. Goodhart-prone.
- **Value model (V)** — a *second* LLM-scale net for PPO's advantage estimation. ~doubles training memory.
- **Reference model (π_ref)** — frozen SFT copy, for the KL constraint. *Third* full model in memory.

$$L^{\mathrm{CLIP}}(\theta) \;=\; \mathbb{E}_t\Big[\min\Big(r_t(\theta)\,\hat{A}_t,\;\mathrm{clip}\big(r_t(\theta),\,1-\epsilon,\,1+\epsilon\big)\,\hat{A}_t\Big)\Big]$$



DPO: Killing the Reward Model

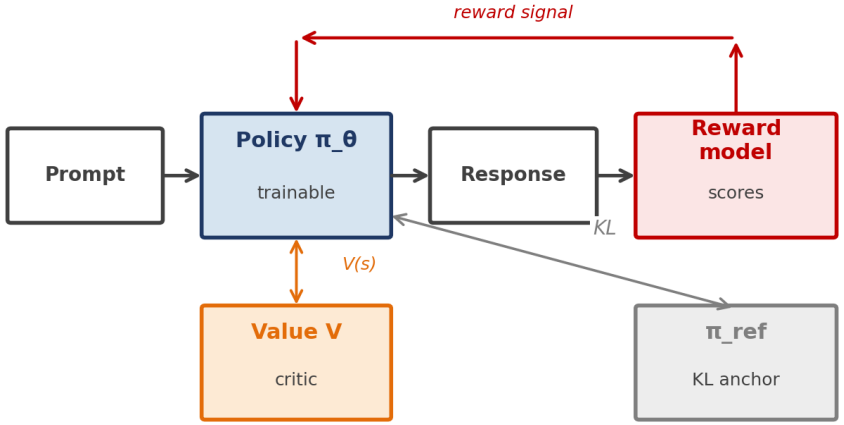
Rafailov et al. Direct Preference Optimization. NeurIPS 2023 (Outstanding Paper)

The insight: under the Bradley-Terry preference model, the *optimal* RLHF policy has a closed form in terms of π_ref and the implicit reward. Invert that — train the *policy* directly to satisfy the preference data, no RM, no RL.

$$\mathcal{L}_{\mathrm{DPO}}(\pi_\theta; \pi_{\mathrm{ref}}) = -\mathbb{E}_{(x,y_w,y_l)}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\mathrm{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\mathrm{ref}}(y_l \mid x)}\right)\right]$$

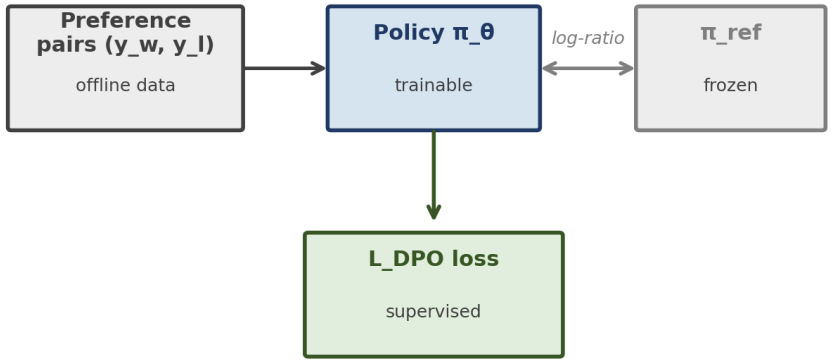
What it removed: the reward model, the value model, and the entire on-policy sampling loop. Training reduces to supervised classification on preference pairs.

PPO-RLHF



4 LLM-scale models · on-policy RL loop

DPO



2 LLM-scale models · supervised, no RL

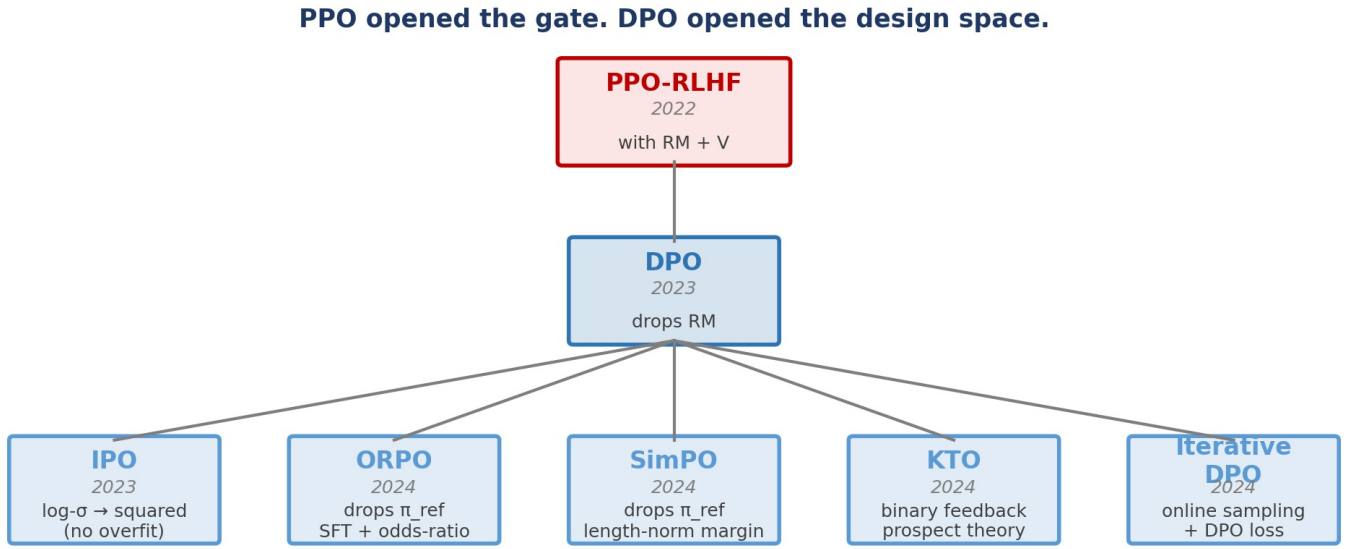
The DPO Variant Explosion

Azar et al. IPO 2023 / Hong et al. ORPO 2024 / Meng et al. SimPO 2024 / Ethayarajh et al. KTO 2024

DPO opened a design space. Each variant fixed one DPO failure mode:

- **IPO** — DPO overfits when preference is deterministic (margin → ∞). Replaces log-sigmoid with a squared loss.
- **ORPO** — drop π_ref entirely; fold preference into the SFT loss via an odds-ratio penalty. *Single-stage* alignment.
- **SimPO** — drop π_ref, replace with a length-normalized margin. Simpler than ORPO, often stronger.
- **KTO** — preferences are expensive; binary thumbs-up/down feedback works almost as well (prospect theory).

Caveat — the leaderboard arms race: *comparisons across variants are confounded by data, base model, and hyperparams. The "best" variant in any paper is usually that paper's variant.*



GRPO: Killing the Value Model

Shao et al. DeepSeekMath. arXiv 2024 / DeepSeek-AI. DeepSeek-R1. Nature 2025

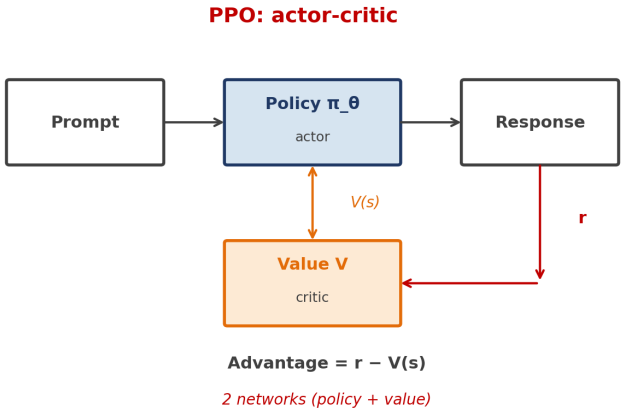
The problem PPO solves with a value model: variance reduction. Advantage = r − V(s) stabilizes the gradient.

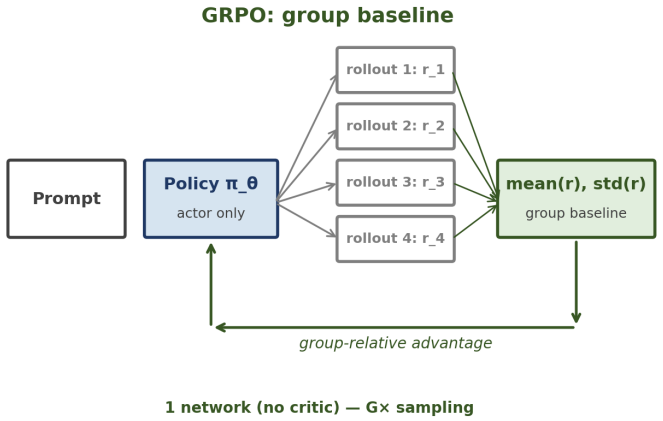
GRPO's substitute: sample G rollouts per prompt; use the *group mean* as the baseline, *group std* as the normalizer. No critic.

$$\hat{A}_{i,t} = \frac{r_i - \text{mean}\big(\{r_1, r_2, \ldots, r_G\}\big)}{\text{std}\big(\{r_1, r_2, \ldots, r_G\}\big)}$$

Cost: G× more sampling per prompt. **Benefit:** ~half the memory; no critic-policy mismatch; clean signal when rewards are sparse/binary (math, code).

Why it caught fire: DeepSeek-R1 (Jan 2025) used GRPO + verifiable rewards to reach o1-level reasoning at a fraction of training cost. Now standard for reasoning-RL.





RLVR: Killing the Human Labels

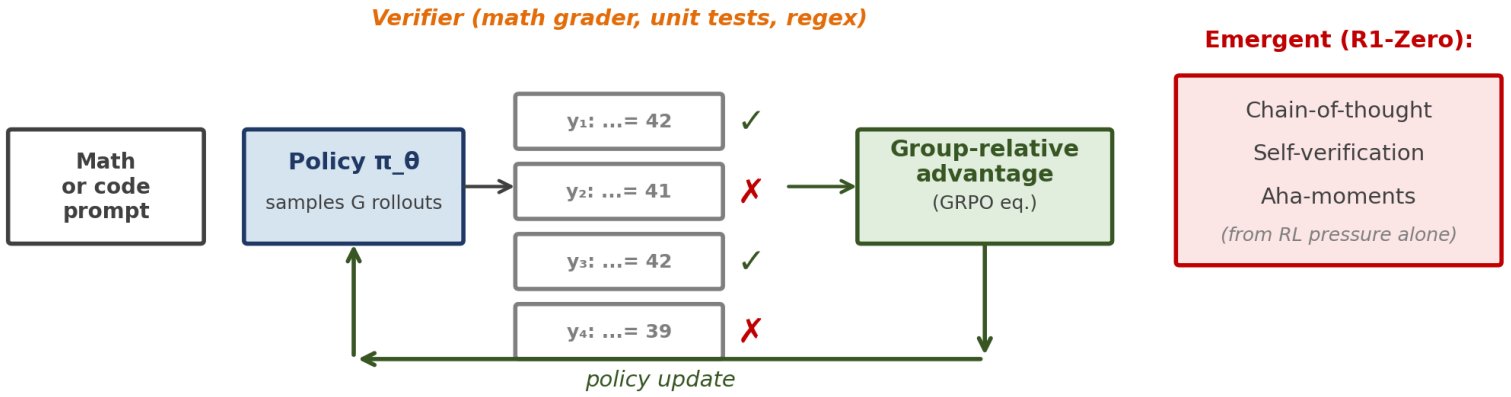
Lambert et al. Tulu 3. arXiv 2024 / Guo et al. DeepSeek-R1. Nature 2025

The problem: human preference labels are expensive, slow, and noisy. RMs trained on them inherit the noise and the Goodhart risk.

RLVR (RL with Verifiable Rewards): replace the RM with a *deterministic verifier*.

- Math: did the boxed answer match ground truth? *(binary, free)*
- Code: did the unit tests pass? *(binary, free)*
- Format: did the output follow the required schema? *(binary, free)*

The R1-Zero result: starting from a base model with *no* SFT, RLVR + GRPO alone produces emergent chain-of-thought, self-verification, and aha-moments.



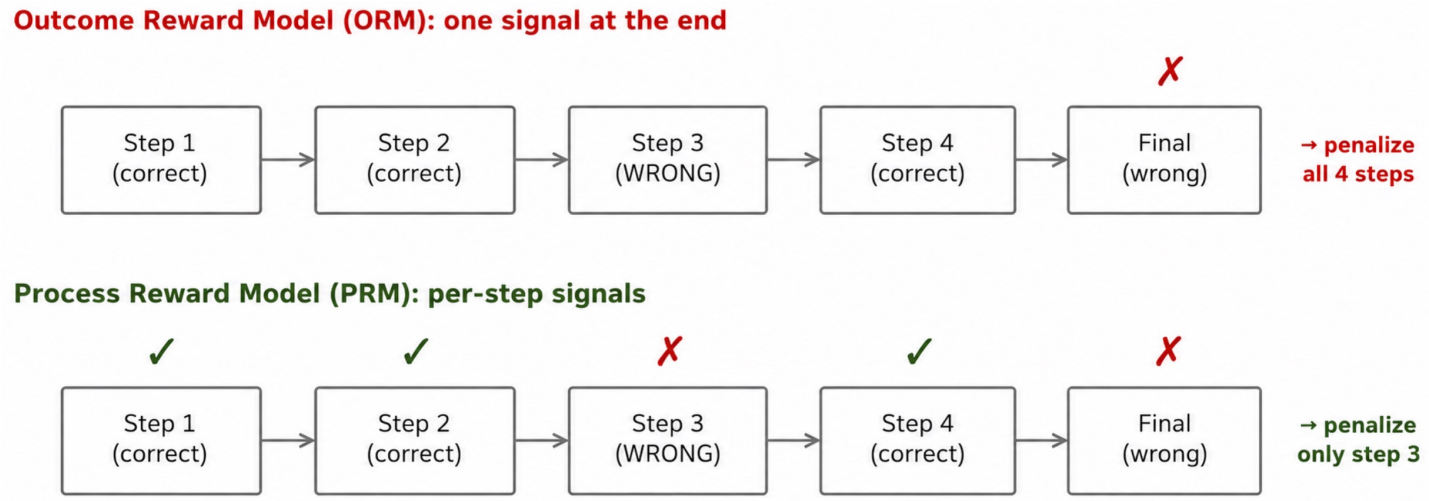
No reward model · No value model · No human labels

Process vs Outcome Rewards

Lightman et al. Let's Verify Step by Step. ICLR 2024 / Wang et al. Math-Shepherd. ACL 2024

Outcome reward (ORM): score the final answer. Cheap and easy to verify — but credits *every* step equally, including wrong steps in a correct rollout, and *none* in a partially-correct rollout.

Process reward (PRM): score *each reasoning step*. Dense supervision, finer credit assignment — but needs per-step labels (expensive) or a learned step-verifier.



Empirical finding (Lightman, OpenAI): PRMs substantially outperform ORMs on MATH at the same compute — but the gap shrinks as the base model gets stronger.

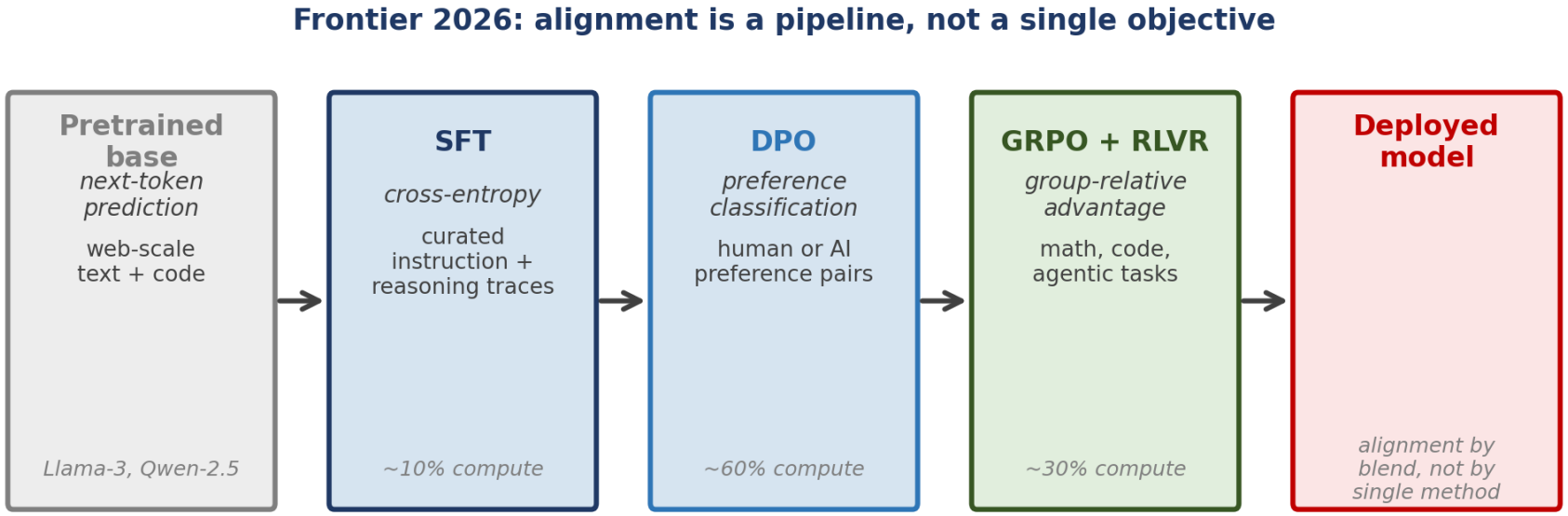
2025 reversal: *DeepSeek-R1 found PRMs hard to scale (reward hacking, label cost). Pure outcome RLVR + GRPO worked better. The PRM-vs-ORM question is still genuinely contested.*

The 2026 Frontier Stack

The post-training recipe at frontier labs (inferred from Llama 3, DeepSeek-R1, Tulu 3, Qwen 2.5):

1. **SFT** on curated instruction data + reasoning traces (distilled or filtered)
2. **DPO** (or variant) on preference data — general helpfulness, harmlessness, style
3. **GRPO + RLVR** on math / code / agentic tasks — verifiable reasoning
4. **Optional**: rejection sampling, iterative DPO, self-play preference generation

The decomposition: *different objectives get different RL. Not one universal method.*

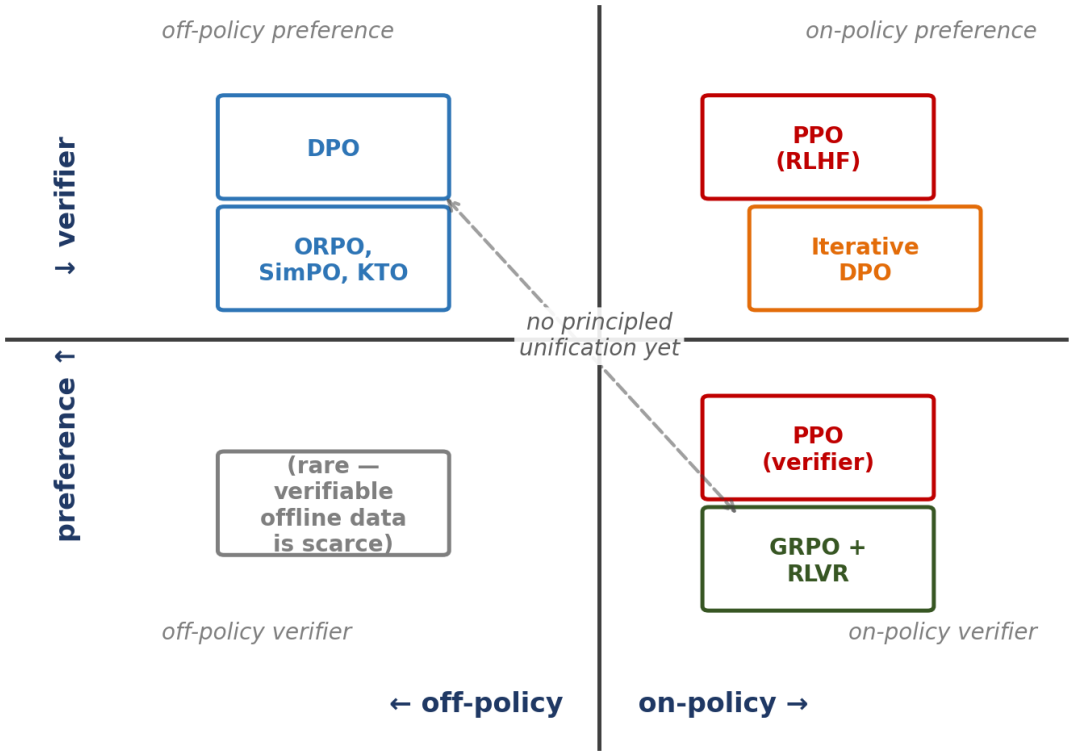


Two Open Problems

(1) Reward hacking is unsolved.
Every successor inherits the original RL problem: optimizing a proxy for what we want. RLVR mostly dodges it (binary verifiers are hard to game) — but only on tasks with verifiers. Outside that, the cat-and-mouse continues.
Gao et al. Scaling Laws for Reward Model Overoptimization (ICML 2023): hacking grows predictably with KL drift.

(2) Off-policy data is a missing piece.
DPO is fully off-policy (a feature). GRPO is fully on-policy (expensive). Frontier labs interleave them, but no principled unification exists yet.
Active: Iterative DPO, Online DPO, REBEL, NashLLM.

Meta-question: is RL even the right frame?
SFT-only or distillation-only pipelines (gemma-2 IT, Phi-3) are competitive on many benchmarks. The case for RL post-training is strongest on *reasoning* and *agentic* tasks — exactly where verifiable rewards exist.



Bridges Back to the Course

To L33 (PPO): every method here either replaces PPO's loss (DPO) or PPO's components (GRPO drops V; RLVR drops RM). The PPO derivation is still the foundation.

To L34 (Model-Based RL): GRPO's group baseline is a *Monte-Carlo* baseline — exactly the variance-reduction trick that motivated TD learning in L31. Same problem, different decade.

To Week 13 (AI Agents): agentic RL is the next frontier — RLVR-on-tool-use, RLAIF for self-improvement, multi-turn credit assignment. The verifier becomes "did the agent complete the task?"

Course thesis, restated: *progress = constraints removed. (W11 said this for games. It holds for post-training too.)*

References

Foundation

[1] Schulman et al. Proximal Policy Optimization Algorithms. arXiv 2017.

[2] Ouyang et al. Training language models to follow instructions with human feedback (InstructGPT). NeurIPS 2022.

DPO family

[3] Rafailov et al. Direct Preference Optimization. NeurIPS 2023.

[4] Azar et al. A General Theoretical Paradigm to Understand Learning from Human Preferences (IPO). arXiv 2023.

[5] Hong et al. ORPO: Monolithic Preference Optimization without Reference Model. EMNLP 2024.

[6] Meng et al. SimPO: Simple Preference Optimization with a Reference-Free Reward. NeurIPS 2024.

[7] Ethayarajh et al. KTO: Model Alignment as Prospect Theoretic Optimization. ICML 2024.

GRPO & RLVR

[8] Shao et al. DeepSeekMath: Pushing the Limits of Mathematical Reasoning. arXiv 2024.

[9] DeepSeek-AI. DeepSeek-R1: Incentivizing Reasoning Capability via Reinforcement Learning. Nature 2025.

[10] Lambert et al. Tülu 3: Pushing Frontiers in Open Language Model Post-Training. arXiv 2024.

Process rewards & open problems

[11] Lightman et al. Let's Verify Step by Step. ICLR 2024.

[12] Wang et al. Math-Shepherd: Verify and Reinforce LLMs Step-by-step. ACL 2024.

[13] Gao et al. Scaling Laws for Reward Model Overoptimization. ICML 2023.