RLVR and RLHF

Pretraind ⟶ Post - trained

Intelligent autocomplete ⟶ Intelligent Assistant.

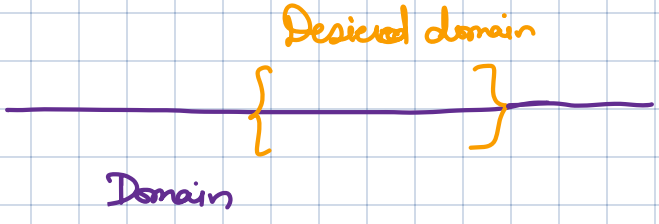
Mixture of training data → autocomplete is a mixture of behaviours

---

We want to amplify good responses, and reduce the bad ones.

Fundamentally, next token prediction is a multiclass classifier
↳ predict probabilities over next token.

→ So this is good if train and test are similar.
But far away ?



→ SGD copies features from training into weights.
↳ So tricky if far away

→ But generated text is not quite like human text

→ Chicken/egg problem → Classifier / model shapes the input
We take a journey steered by what we can generate.

Examples

- Good examples : GD
- Bad —"— : GA

→ ① Ask people - please label for me
↳ very hard

② Verifiable rewards → 1) Math problems with numerical ans
2) Proofs
3) Software w/ test cases

③ AI feedback → let us have model that is trained for this

→ Parallel rollouts - try solving problem many times.

→ Grade each one

→ good examples → bad examples

→ How to generate?

→ What do we do with the labels

Data Generation : Chain-of-Thought.

Reinforce

→ If you have something non-differentiable, think of it as diff.
→ can you make it differentiable

Reward: $y \to F(y)$ $\quad F(y) = \begin{cases} 0 & \text{if wrong} \\ 1 & \text{if correct} \end{cases}$

$$J(\theta) = E_{y \sim P(\theta)} [F(y)]$$

θ: model parameters.

$y \sim P(\theta)$: differentiable function of θ

$$E_{y \sim P(\theta)} [F(y)] = \sum_y P_\theta(y) \cdot F(y)$$

$$\nabla_\theta E [F(y)] = \sum_y \nabla_\theta (P_\theta(y) \cdot F(y))$$

$$= \sum_y F(y) \nabla_\theta (P_\theta(y)).$$

Trick: $\frac{d}{dt} \ln(f(t)) = \frac{1}{f(t)} \cdot \frac{d}{dt} f(t)$.

$$= \sum_y P_\theta(y) \cdot F(y) \cdot \frac{1}{P_\theta(y)} \nabla_\theta (P_\theta(y))$$

prob. → $P_\theta(y)$; term ← $F(y) \nabla_\theta (\ln P_\theta(y))$

$$= \sum_y P_\theta(y) \left[F(y) \nabla_\theta (\ln P_\theta(y)) \right]$$

$$= E_{y \sim P(\theta)} \left[F(y) \nabla_\theta (\ln P_\theta(y)) \right]$$

RLVR : Reinforcement Learning Verifiable Rewards.

Scale RL Paper - Khatri et al. 2025.

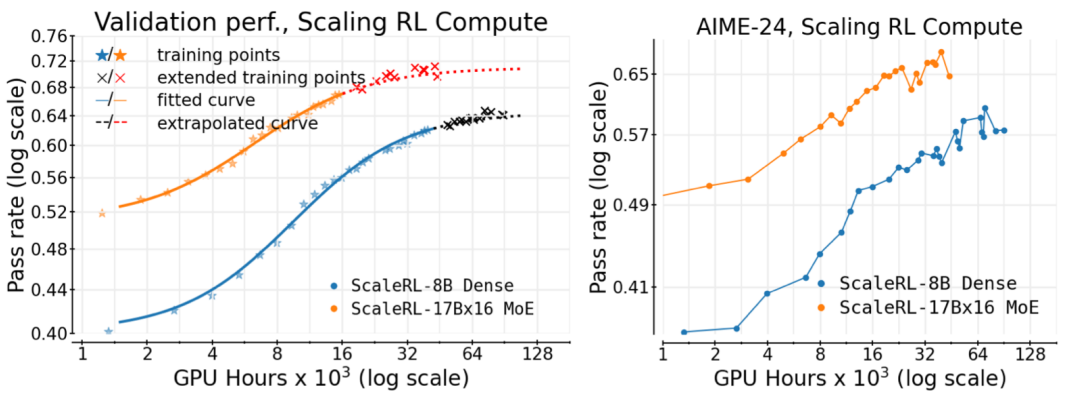


Fig 1

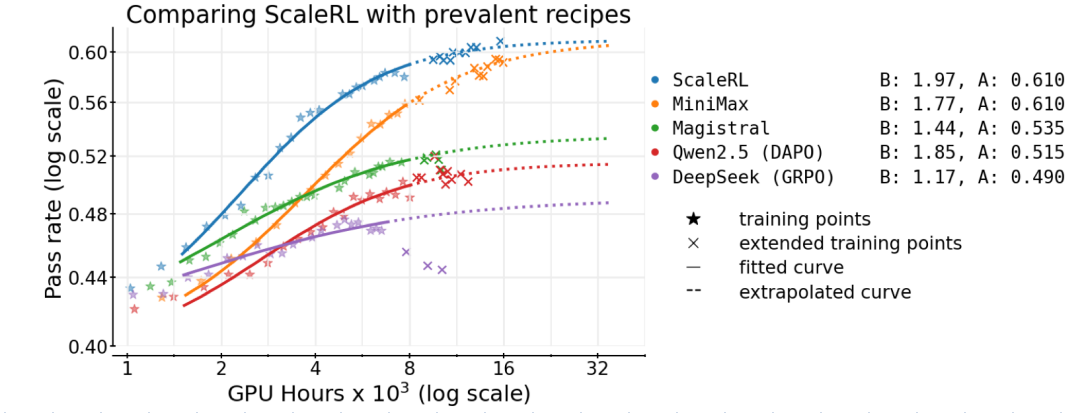


Fig 2

$$\mathcal{J}_{\text{ScaleRL}}(\theta) = \mathop{\mathbb{E}}_{\substack{x\sim D,\\ \{y_i\}_{i=1}^{G}\sim\pi_{gen}^{\theta_{old}}(\cdot|x)}} \left[\frac{1}{\sum_{g=1}^{G}|y_g|} \sum_{i=1}^{G}\sum_{t=1}^{|y_i|} \text{sg}(\min(\rho_{i,t}, \epsilon))\hat{A}_i^{\text{norm}} \log \pi_{train}^{\theta}(y_{i,t}) \right],$$

$$\rho_{i,t} = \frac{\pi_{train}^{\theta}(y_{i,t})}{\pi_{gen}^{\theta_{old}}(y_{i,t})}, \quad \hat{A}_i^{\text{norm}} = \hat{A}_i/\hat{A}_{\text{std}}, \quad 0 < \text{mean}(\{r_j\}_{j=1}^{G}) < 1, \quad \text{pass\_rate}(x) < 0.9,$$

log prob

only look at imperfect prompts

G: # of responses generated

Generation may be done by a model different from the one being trained

↳ code may be optimized differently.

y_i: ith response

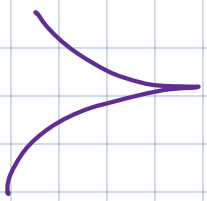
sg: Stop gradient

r_i: Grade of response i 0 or 1.

$\bar{r}(x)$: $\frac{1}{G}\sum_j r_j$

$\hat{A}_i = r_i - \bar{r}(x)$ advantage of i

RLHF



① Train a reward model on comparisons. (x, y_w, y_l)

y_w is preferred response

$$P(y_w \succ y_l \mid x) = \sigma(r(x, y_w) - r(x, y_l))$$

Bradley-Terry model.

Train a reward model to maximize likelihood of observed preference

$$\mathcal{L}_{RM}(\phi) = -\mathbb{E}_{(x, y_w, y_l)}\left[\log \sigma(r_\phi(x, y_w) - r_\phi(x, y_l))\right]$$

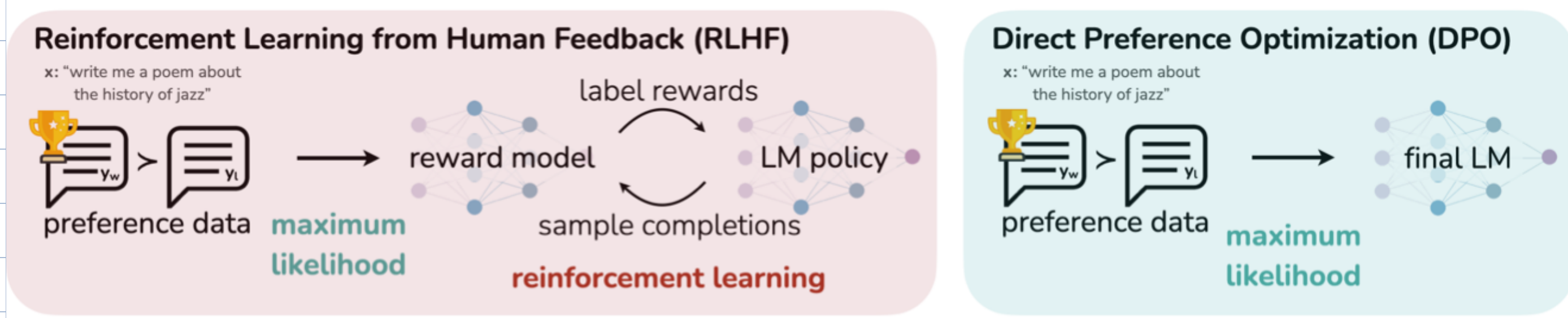
② Next RL

$$\max_{\pi} \mathbb{E}_x \left[\mathbb{E}_{y \sim \pi(\cdot|x)}[r_\phi(x, y)] - \beta D_{KL}(\pi(\cdot|x) \,||\, \pi_{ref}(\cdot|x))\right]$$

DPO : Direct Preference Optimization

Can we train on (x, y_w, y_l) without a reward model?

And no RL loop?



① Idea 1

$$\max \log \pi_\theta(y_w) - \log \pi_\theta(y_l)$$

Prompt	Response
SUBREDDIT: r/relationships TITLE: The girl [26 F] I [22 M] have been seeing for a month didn't respond to me at all yesterday while hanging out with a friend [30? M]. POST: She gets terrible service while at her house, but I texted her 3 times yesterday, 4-5 hours apart. She didn't call me until early this morning and left a voicemail that she was busy all day with a friend who showed up out of the blue. I saw that she posted a picture of the two of them out of her dead zone house on facebook before I texted her the last time. I don't mind that she hangs out with friends, and I know it's pretty early [...] TL;DR:	girl whenUB when when when when whenUB when when when whenUB
SUBREDDIT: r/tifu TITLE: TIFU by accidently kicking an old woman POST: So this didn't happen today but actually about a year or two ago. I was at my granddads funeral so of course it was all very sad and full of lots of crying old people. After the ceremony everyone walks outside the building and onto the other side of the small road the hearses drive down. Now the road is important because obviously if there's a road, there's a curb onto the sidewalk, so most of us are on the other side of the road, besides a few older people walking a lot slower. As one of the old woman goes to walk up the curb [...] TL;DR:	when an old woman was tripping the when

(2) SFT on winners

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

$$\nabla_\theta\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) =$$
$$-\beta\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Bigg[\underbrace{\sigma(\hat{r}_\theta(x,y_l) - \hat{r}_\theta(x,y_w))}_{\text{higher weight when reward estimate is wrong}}\Bigg[\underbrace{\nabla_\theta\log\pi(y_w \mid x)}_{\text{increase likelihood of } y_w} - \underbrace{\nabla_\theta\log\pi(y_l \mid x)}_{\text{decrease likelihood of } y_l}\Bigg]\Bigg],$$

where $\hat{r}_\theta(x,y) = \beta\log\frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$