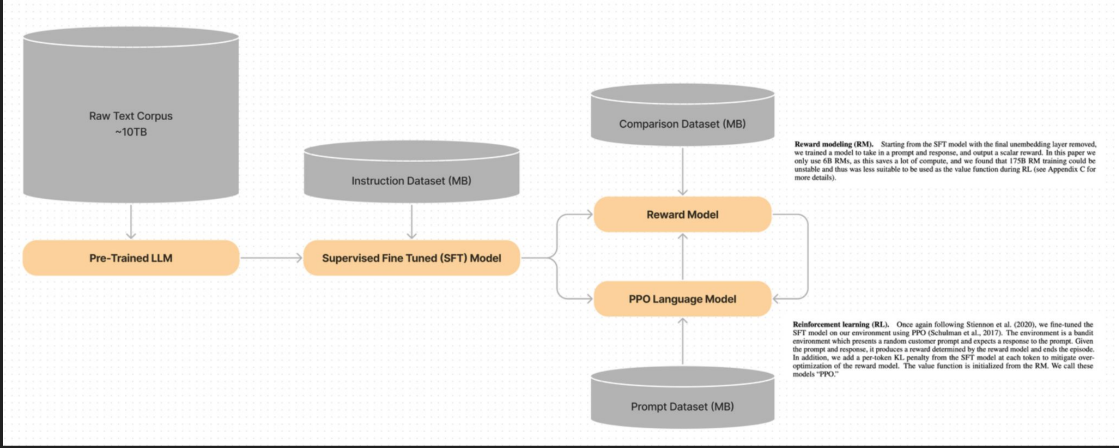
RLHF and Its Alternatives

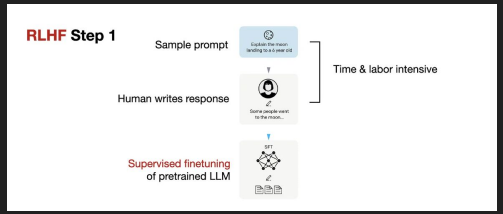
Andrey Arzhantsev

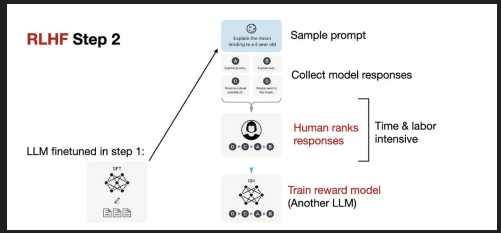
Plan

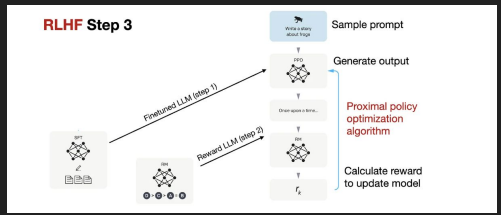
- RLHF variations
  - ChatGPT RLHF Setup (OpenAI)
  - Llama2 RLHF Setup (Meta)
  - Sparrow RLHF Setup (DeepMind)
  - Other RLHF Setups
- RLHF alternatives
  - RLHF Disadvantages (RL and HF)
  - Contrastive Preference Learning (Stanford University & Smth.)
  - DPO: Your LM is Secretly a RM (Stanford University)
  - ReST (DeepMind)
  - AIF over HF
    - RLAIF (Google Research)
    - Red Teaming; Constitutional AI (Anthropic)

ChatGPT RLHF Setup (OpenAI)



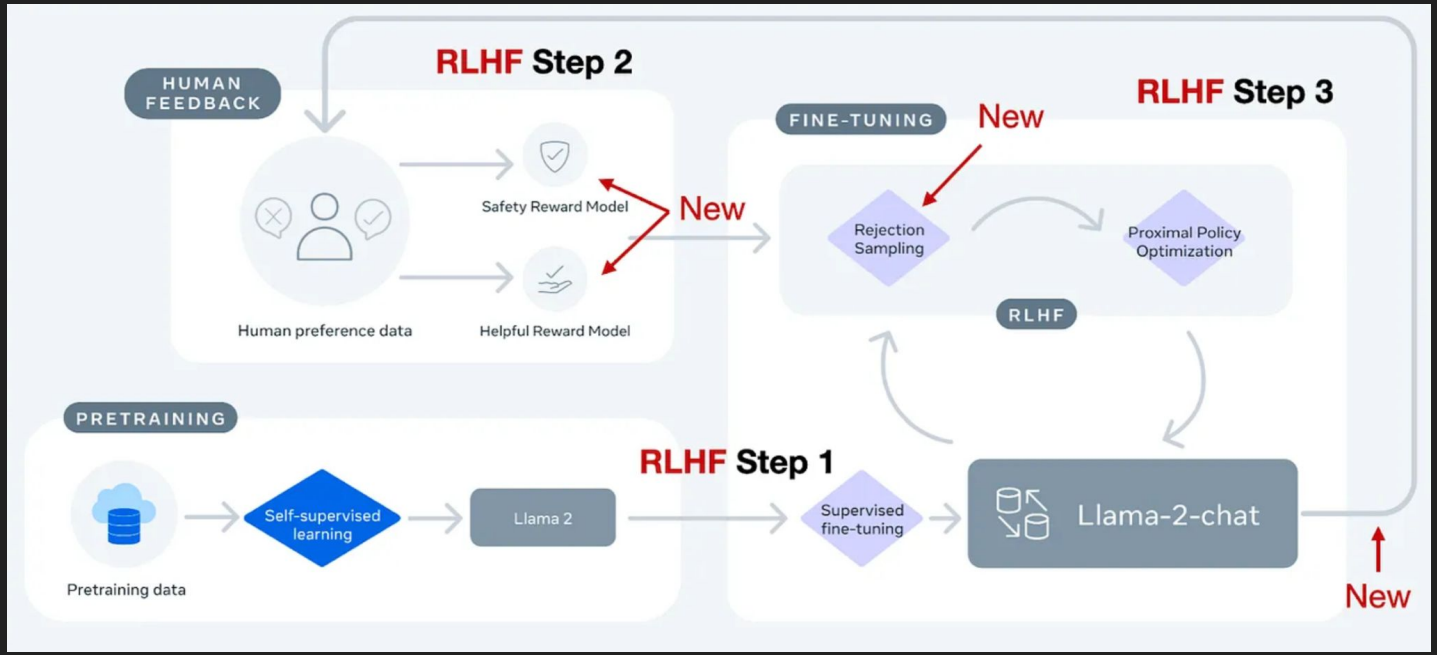






Instruct GPT Paper:
https://arxiv.org/abs/2203.02155
My presentation a year ago:
https://docs.google.com/presentation/d/1yqfF5Z5Vg2JYM1_UW_dwLMOFtibQL_kvuuP5O3Mg1XM/edit?usp=sharing

Llama2 RLHF Setup (Meta)

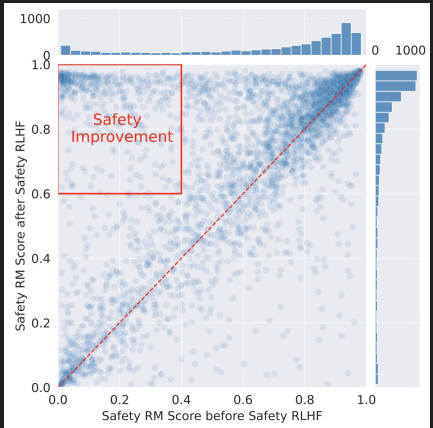


Llama 2 Paper:
https://arxiv.org/pdf/2307.09288.pdf

Llama2 RLHF Setup (Meta)

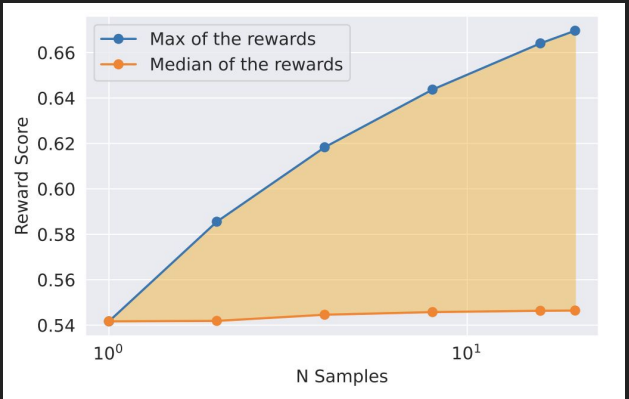
Novelties:

1. Two reward models



benefits from Safety RM

2. Rejection sampling



Explanation of benefit of Rejection Sampling

3. Margin loss

$$\mathcal{L}_{\text{ranking}} = -\log(\sigma(r_\theta(x, y_c) - r_\theta(x, y_r) - m(r)))$$

	Significantly Better	Better	Slightly Better	Negligibly Better / Unsure
Margin Small	1	2/3	1/3	0
Margin Large	3	2	1	0

Table 27: Two variants of preference rating based margin with different magnitude.

	Significantly Better	Better	Slightly Better	Negligibly Better / Unsure	Avg
No margin	79.1	66.9	59.8	54.5	62.5
Margin Small	80.4	67.3	60.4	**55.0**	**63.0**
Margin Large	**80.7**	**67.5**	**60.5**	54.3	62.9

Table 28: Ablation on preference rating-based margin in Helpful reward model ranking loss. The rating margin component helps improve model accuracy on samples with more separable response pairs (e.g., chosen response significantly better the rejected counterpart).

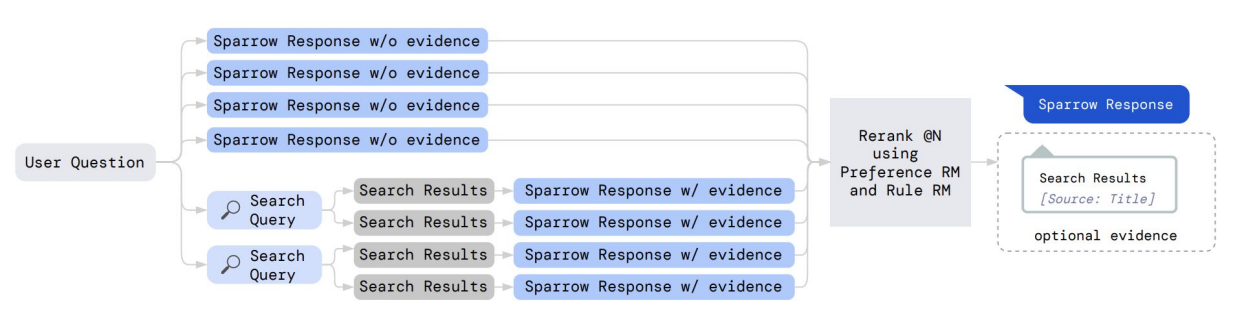
Reweighting labels and loss with margin

Sparrow RLHF Setup (DeepMind)

Novelties:

1. Evidence

Use Google API to generate text from Google Search results with links to provide evidence



User: What color is the sky?

Sparrow: Blue.

User: Why?

Search Query: why is sky blue

Search Results:

Page title: Sky Except for direct sunlight, most of the light in the daytime sky is caused by scattering, which is dominated by a small-particle limit called Rayleigh scattering. The scattering due to molecule-sized particles (as in air) is greater in the directions both toward and away from the source of light than it is in directions perpendicular to the incident path. Scattering is significant for light at all visible wavelengths, but is stronger at the shorter (bluer) end of the visible spectrum, meaning that the scattered light is bluer than its source: the Sun. The remaining direct sunlight, having lost some of its shorter-wavelength components, appears slightly less blue.

Sparrow: Because of the Rayleigh scattering. It causes short waves of blue light to be scattered more than the other colours in the spectrum.

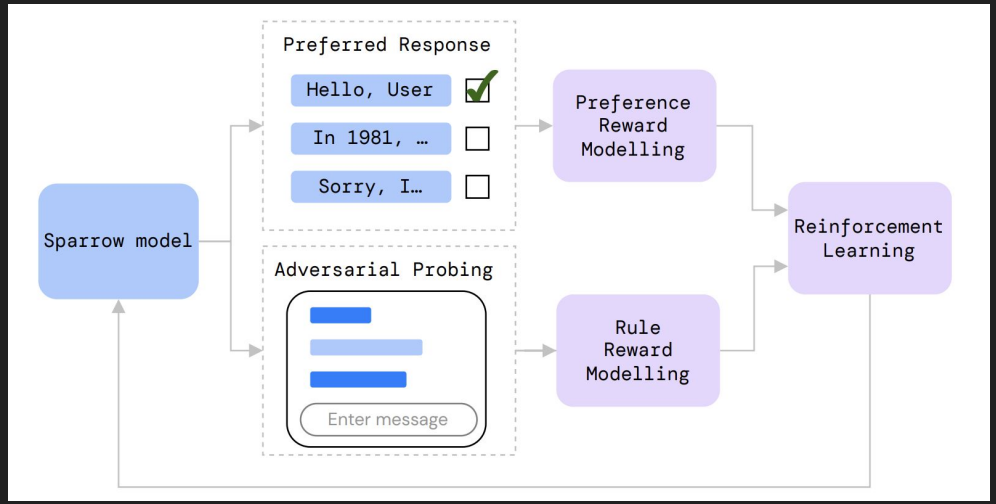
Sparrow 2 Paper:
https://storage.googleapis.com/deepmind-media/DeepMind.com/Authors-Notes/sparrow/sparrow-final.pdf

Sparrow RLHF Setup (DeepMind)

Novelties:

2. Rule Reward Modelling

Use conditional classifier that estimates the probability that the rule X was violated by Sparrow at any point in the dialogue

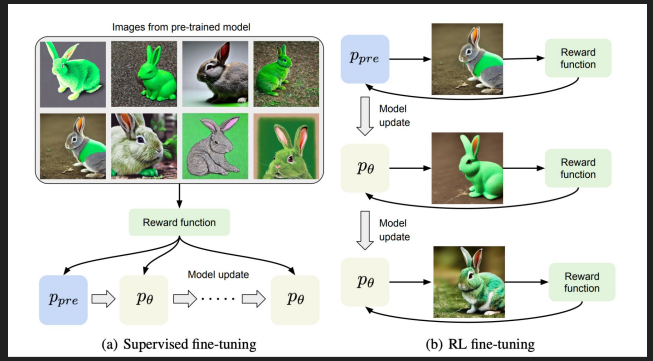


$$R_{\text{rerank}} = \frac{e^{R_{pr}}}{e^{R_{pr}} + e^{AVG(R_{pr})}} \left(\prod_{i=1}^{n} R_{\text{rule}_i} \right)^{\frac{1}{n}}$$

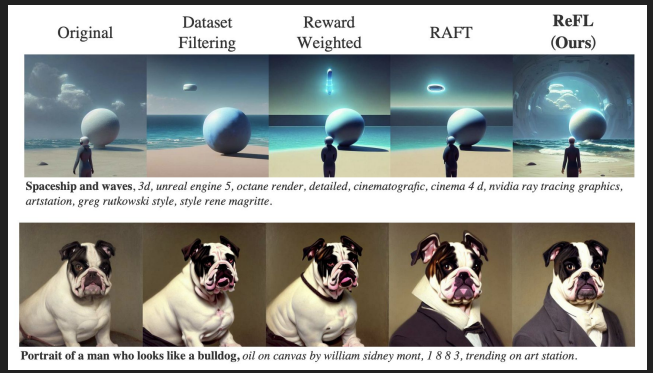
Sparrow 2 Paper:
https://storage.googleapis.com/deepmind-media/DeepMind.com/Authors-Notes/sparrow/sparrow-final.pdf

Other RLHF Setups

- NLP:
  - Gemini (Google)
  - Claude (Anthropic)
- CV:
  - DPOK - Diffusion Policy Optimization with KL regularization (Google, UC Berkeley, etc.)
  - ReFL - Reward Feedback Learning (Tsinghua University, etc.)
- Video game bots:
  - Deep reinforcement learning from human preferences (DeepMind & OpenAI)

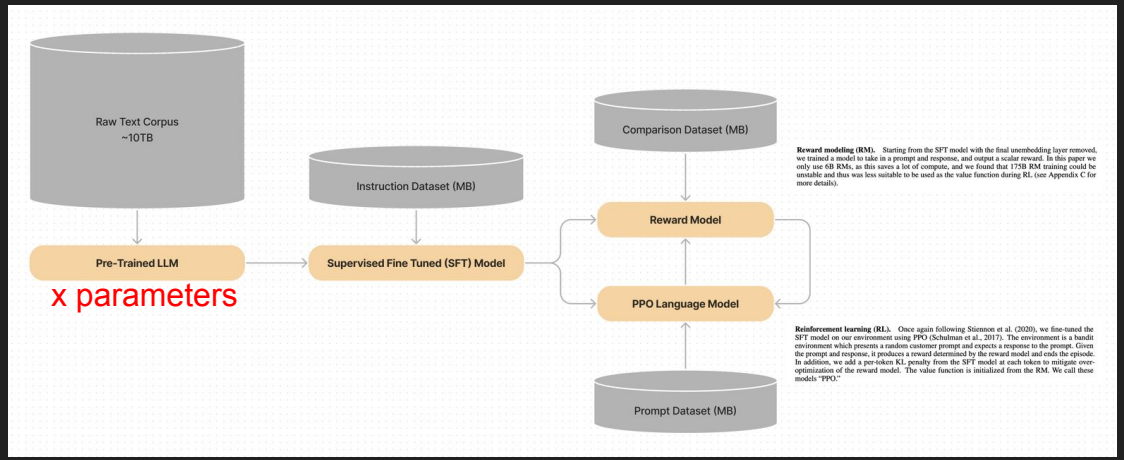


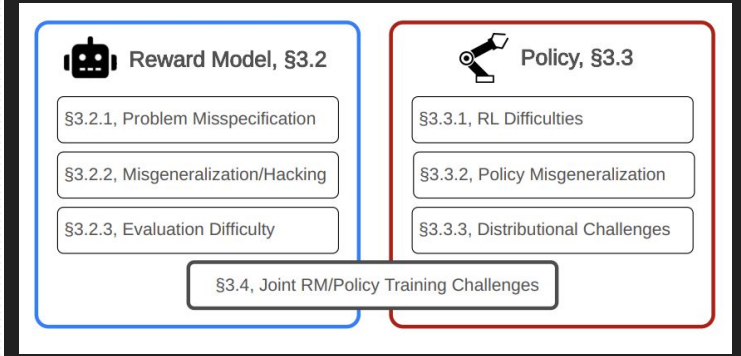
(DPOK)



(ReFL)

RLHF disadvantages №1 - RL



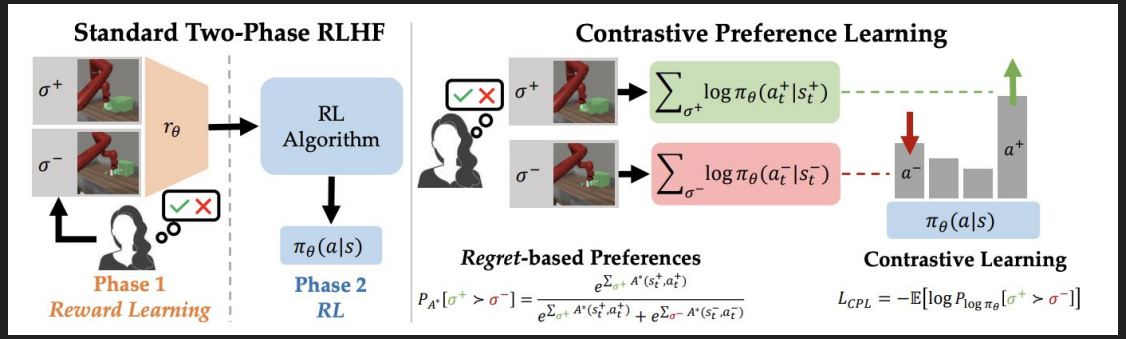


1) Complex and often unstable

2) Each step copy model + train smth => 4 steps = 4x parameters => ~4x memory, more training time, resources, etc.

Contrastive Preference Learning (Stanford University & Smth.)

- Presented for robotics environment and LLM fine tuning
- One step instead of two



CPL Paper:
https://arxiv.org/pdf/2310.13639.pdf

Contrastive Preference Learning

$$\sigma = (s_1, a_1, s_2, a_2, \ldots, s_k, a_k)$$

(Model parameters)

$$\pi^*(a|s) = e^{A_r^*(s,a)/\alpha}$$

(Theorem)

This means that instead of learning the optimal advantage function, we can directly learn the optimal policy

$$V^\pi(s) = \mathop{\mathrm{E}}_{\substack{a\sim\pi \\ s'\sim P}}\left[r(s,a) + \gamma V^\pi(s')\right],$$

$$Q^\pi(s,a) = \mathop{\mathrm{E}}_{s'\sim P}\left[r(s,a) + \gamma \mathop{\mathrm{E}}_{a'\sim\pi}\left[Q^\pi(s',a')\right]\right]$$

$$A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$$

(Useful RL terms)

$$P_{A^*}\left[\sigma^+ \succ \sigma^-\right] = \frac{\exp \sum_{\sigma^+} \gamma^t \alpha \log \pi^*(a_t^+|s_t^+)}{\exp \sum_{\sigma^+} \gamma^t \alpha \log \pi^*(a_t^+|s_t^+) + \exp \sum_{\sigma^-} \gamma^t \alpha \log \pi^*(a_t^-|s_t^-)}$$

(Consequence)

$$\sum_{t=1}^{k} \gamma^t r_E(s_t, a_t)$$

$$-\sum_{t=1}^{k} \gamma^t (V^*(s_t) - Q^*(s_t, a_t))$$

(Preference model goal change)

$$\mathcal{L}_{\mathrm{CPL}}(\pi_\theta, \mathcal{D}_{\mathrm{pref}}) = \mathbb{E}_{(\sigma^+,\sigma^-)\sim\mathcal{D}_{\mathrm{pref}}}\left[-\log \frac{\exp \sum_{\sigma^+} \gamma^t \alpha \log \pi_\theta(a_t^+|s_t^+)}{\exp \sum_{\sigma^+} \gamma^t \alpha \log \pi_\theta(a_t^+|s_t^+) + \exp \sum_{\sigma^-} \gamma^t \alpha \log \pi_\theta(a_t^-|s_t^-)}\right]$$

(Final loss function)

DPO: Your LM is Secretly a RM (Stanford University)

- It is actually a special case of CPL:
  - MDP terminates after a single step and there is no next state

$$A^*(s,a) = Q^*(s,a) - V^*(s,a) = r_E(s,a) + \gamma \mathbb{E}_{s'}[V^*(s')] - V^*(s)$$

The regret preference model becomes:

$$P_{A^*}\left[\sigma^+ \succ \sigma^-\right] = \frac{\exp r_E(s,a^+) - V^*(s)}{\exp r_E(s,a^+) - V^*(s) + \exp r_E(s,a^-) - V^*(s)} = \frac{\exp r_E(s,a^+)}{\exp r_E(s,a^+) + \exp r_E(s,a^-)}$$

$$\mathcal{L}_{\text{CPL}}(\pi_\theta, \mathcal{D}_{\text{pref}}) = \mathbb{E}_{(\sigma^+,\sigma^-)\sim\mathcal{D}_{\text{pref}}}\left[-\log \frac{\exp \sum_{\sigma^+} \gamma^t \alpha \log \frac{\pi_\theta(a_t^+|s_t^+)}{\mu(a_t^+|s_t^+)}}{\exp \sum_{\sigma^+} \gamma^t \alpha \log \frac{\pi_\theta(a_t^+|s_t^+)}{\mu(a_t^+|s_t^+)} + \exp \sum_{\sigma^-} \gamma^t \alpha \log \frac{\pi_\theta(a_t^-|s_t^-)}{\mu(a_t^-|s_t^-)}}\right]$$

DPO Paper:
https://arxiv.org/abs/2305.18290

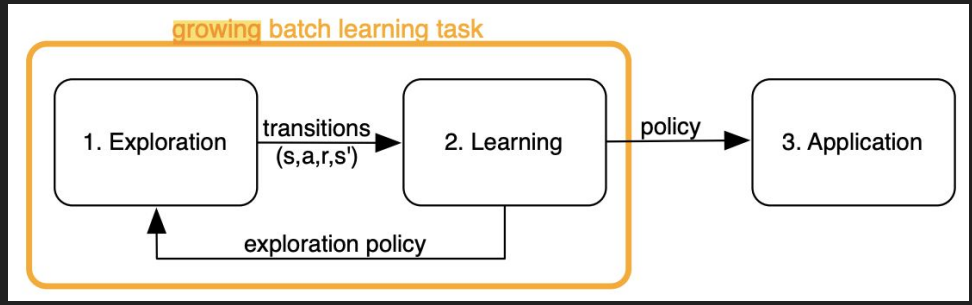
ReST (DeepMind)

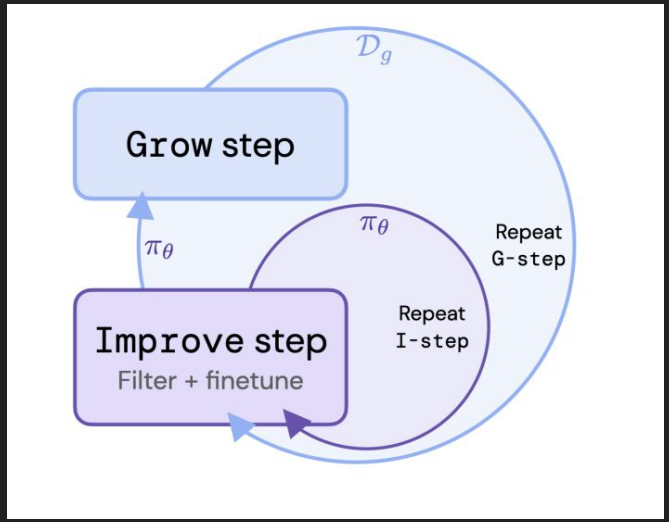
- Setup: growing batch RL problem
- 2 steps:
  - Grow step: 1) policy is used to generate predictions for each context to augment the training dataset 2) Annotate new dataset with the reward model $R(x, y)$
  - Improve step: 1) rank and filter dataset from Grow step 2) finetune model on it using offline RL objective

$$F(x, y; \tau) = \mathbb{1}_{R(x,y)>\tau} \cdot$$

$$\mathcal{L}_{\text{NLL}}(\theta) = -\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\sum_{t=1}^{T} \log \pi_\theta(y_t \mid y_{1:t-1}, x)\right]$$

$$J(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}_g}\left[F(x, y; \tau)\ \mathcal{L}(x, y; \theta)\right]$$





ReST Paper:
https://arxiv.org/pdf/2308.08998.pdf

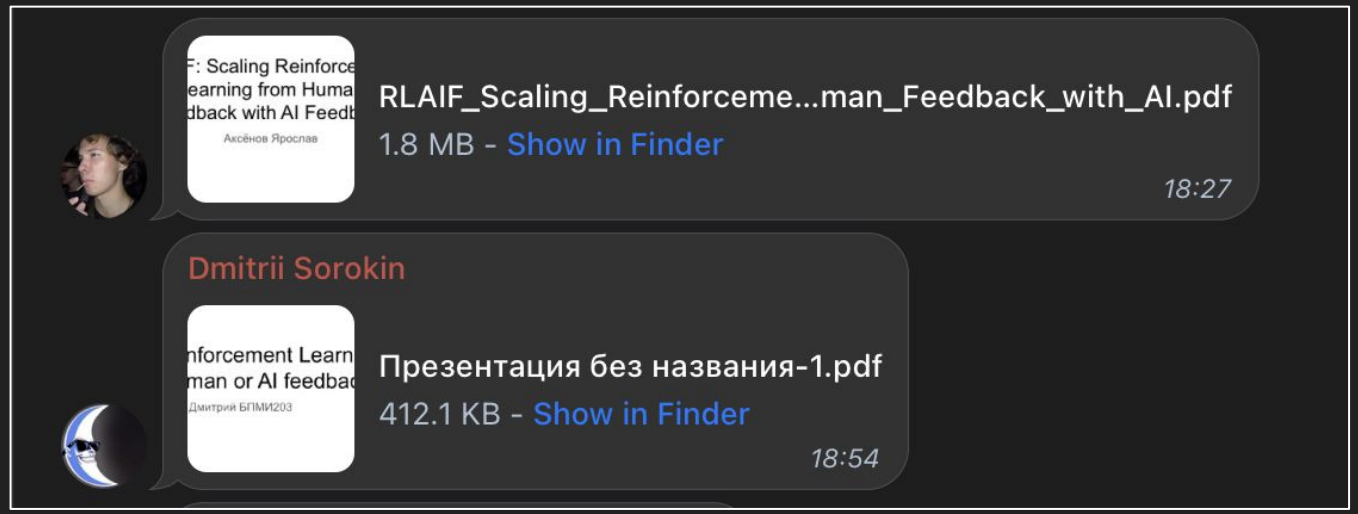
RLHF disadvantages №2 - HF

Human Feedback, §3.1

- §3.1.1, Misaligned Evaluators
- §3.1.2, Difficulty of Oversight
- §3.1.3, Data Qualilty
- §3.1.4, Feedback Type Limitations

RLAIF (Google Research)

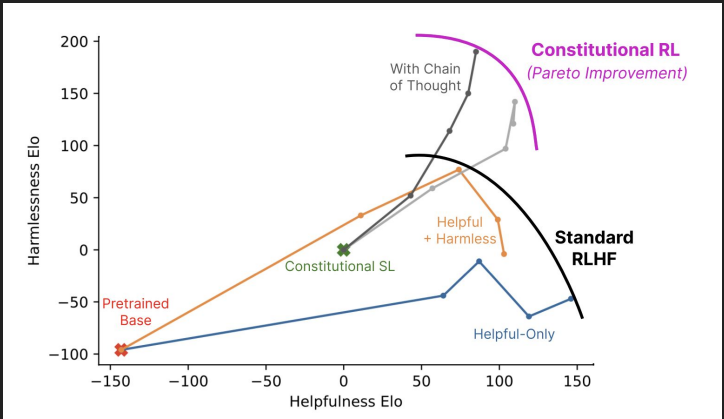
If you missed

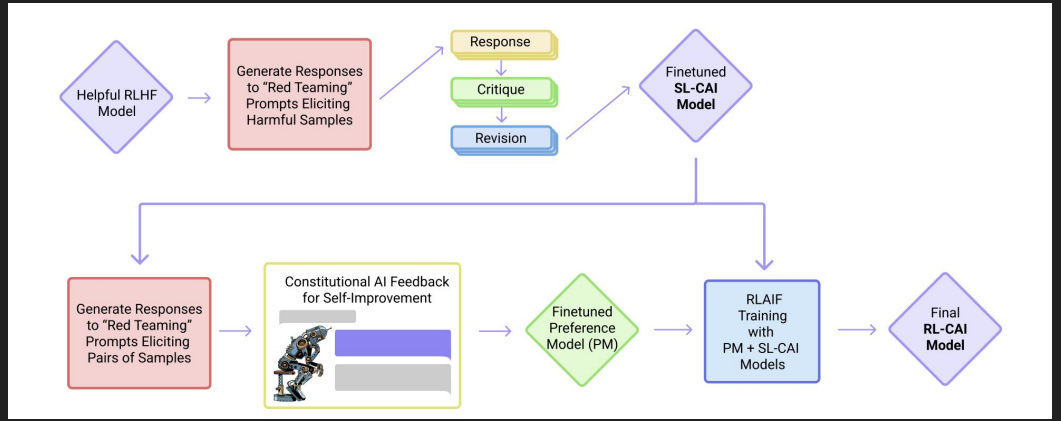


RLAIF Paper:
https://arxiv.org/pdf/2309.00267.pdf

Red Teaming; Constitutional AI (Anthropic)

Red Teaming - an “enemy” that challenges model (it’s harmfulness in this case)





Constitutional AI Paper:
https://arxiv.org/abs/2212.08073

That’s all