Higher School of Economics
Faculty of Computer science

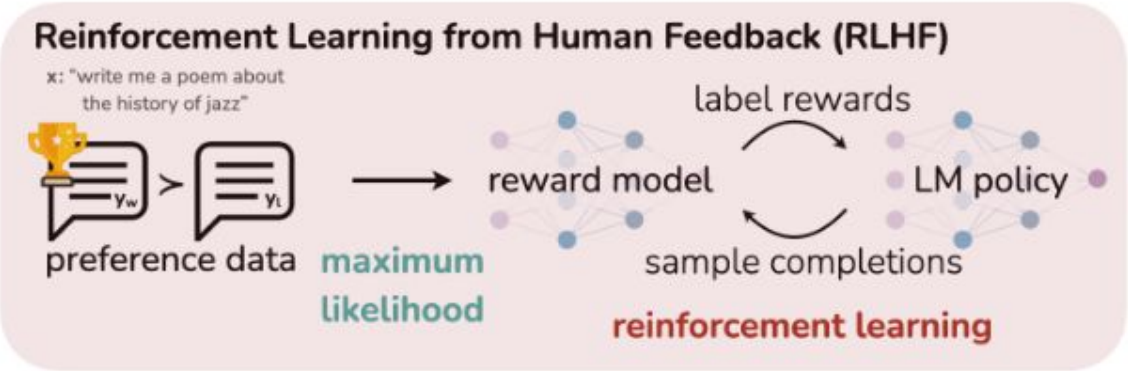
RLHF without RL

Бугаев Егор, ПМИ 213 (НИС МОП)

Что такое RLHF

- Хотим: научить модель выдавать предсказания, лучше соответствующие желаниям людей
- Необходимо: включить людей в процесс обучения

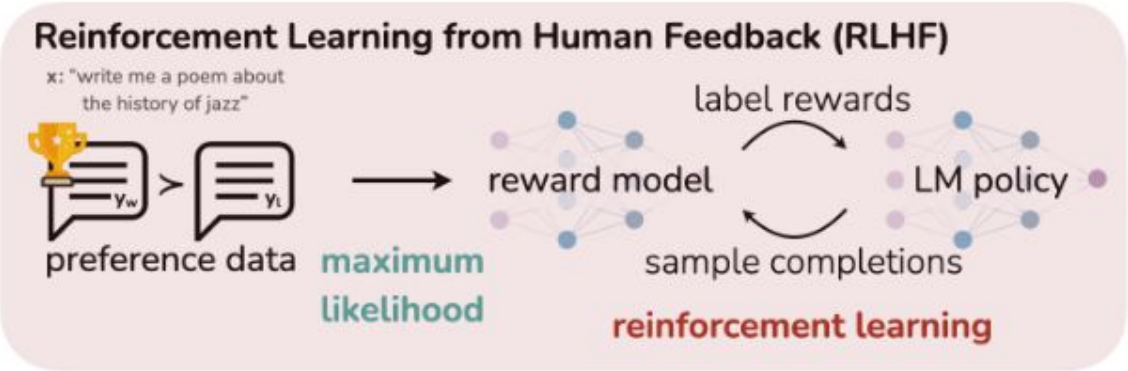
Решение: Reinforcement Learning with Human Feedback



Pipeline для RLHF

Higher School of Economics
Faculty of Computer science

1. Берем обученную модель (далее SFT от Supervised Fine-Tuning)
2. Выбираем набор промптов (x), получаем от модели два предсказания. С помощью людей выбираем y_w, y_l, обучаем на этом reward model.
3. Дообучаем модель с помощью Reinforcement Learning, используя reward model



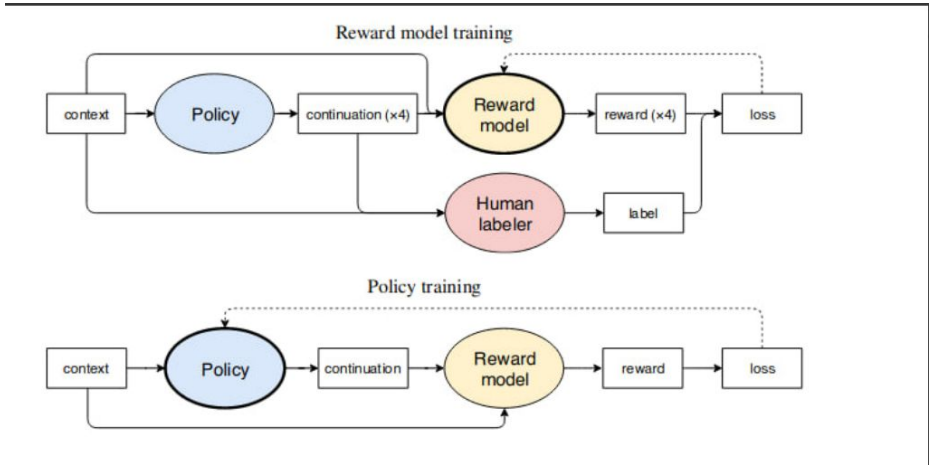
Что за RL в RLHF

Higher School of Economics
Faculty of Computer science

Считаем, что существует скрытая r(x, y) - reward функция

Хотим: reward model(x, y) = r(x, y)

1. Обучаем reward model: генерируем на каждый промпт два ответа, человек оценивает, какой лучше. Обучаем модель как:



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

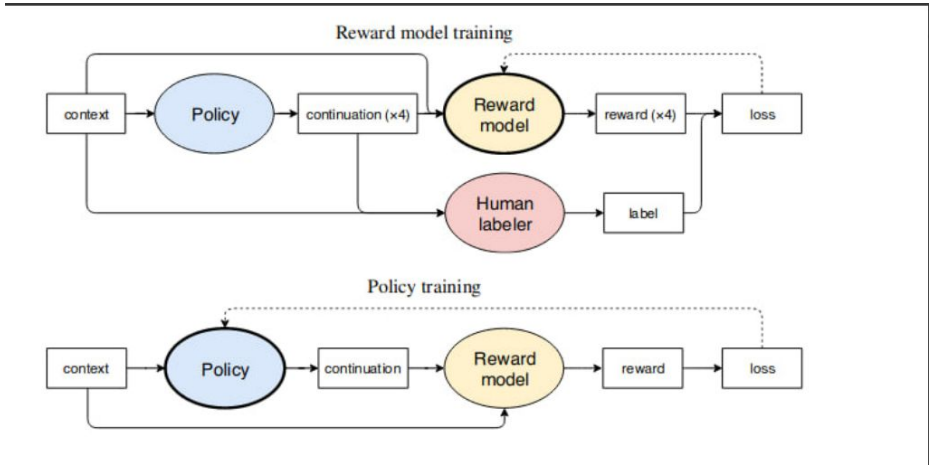
Фактически: хотим максимизировать r_model(x, y_w) - r_model(x, y_l)

Что за RL в RLHF

Higher School of Economics
Faculty of Computer science

2. Тюним SFT модель с помощью полученной r_model:

- Максимизируем награду генерируемых ответов
- При этом хотим не сильно отклониться от исходной модели (аналог регуляризации)



$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(y|x)} \left[r_\phi(x, y) \right] - \beta \mathbb{D}_{\mathrm{KL}} \left[\pi_\theta(y \mid x) \,||\, \pi_{\mathrm{ref}}(y \mid x) \right]$$

А зачем убирать RL?

Higher School of Economics
Faculty of Computer science

Плюсы:

- Большая гибкость метода (сами выбираем reward model)
- Хорошие результаты

Минусы:

- Из-за RL: сложно настроить процесс обучения
- Большие вычислительные затраты (теперь обучаем две модели)

DPO: избавляемся от reward model

Higher School of Economics
Faculty of Computer science

Заметим, что можем выразить оптимальную политику (вероятность) через награду:

$$\pi_r(y \mid x) = \frac{1}{Z(x)} \pi_{\text{ref}}(y \mid x) \exp\left(\frac{1}{\beta} r(x, y)\right),$$

where $Z(x) = \sum_y \pi_{\text{ref}}(y \mid x) \exp\left(\frac{1}{\beta} r(x, y)\right)$ is the partition function.

А теперь выражаем отсюда r(x, y) через политику и подставляем в следующую вероятность (работает для любой пары политика/reward):

$$p^*(y_1 \succ y_2 \mid x) = \frac{1}{1 + \exp\left(\beta \log \frac{\pi^*(y_2|x)}{\pi_{\text{ref}}(y_2|x)} - \beta \log \frac{\pi^*(y_1|x)}{\pi_{\text{ref}}(y_1|x)}\right)}$$

DPO: избавляемся от reward model

Higher School of Economics
Faculty of Computer science

Но тогда мы можем посчитать функцию потерь без reward model:

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

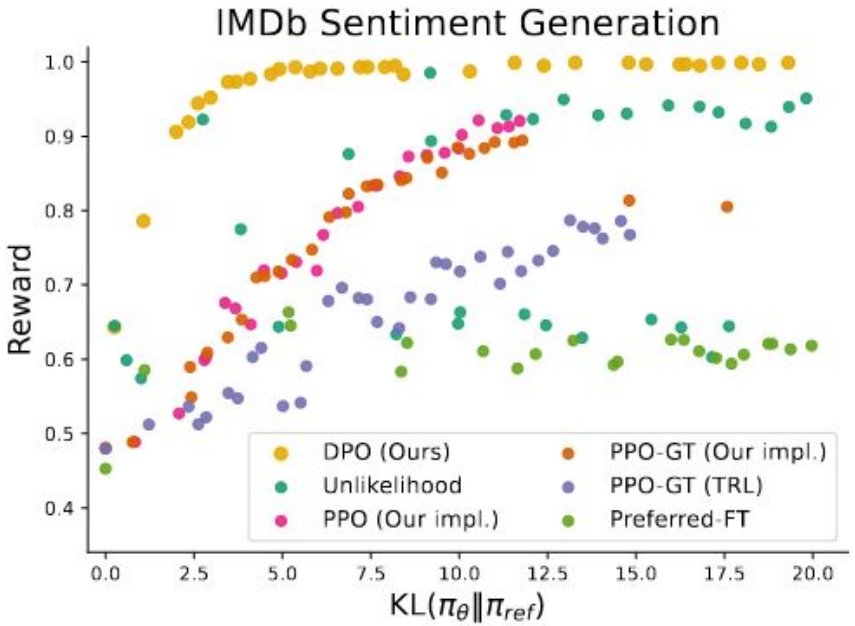
Тогда новый pipeline:

1. На каждый промпт генерируем y_1, y_2. Человек выбирает, какой лучше
2. Отдаем такой размеченный датасет, оптимизируя функцию потерь сверху (теперь просто supervised дообучение, только с двумя x вместо одного)

DPO: эксперименты

Higher School of Economics
Faculty of Computer Science

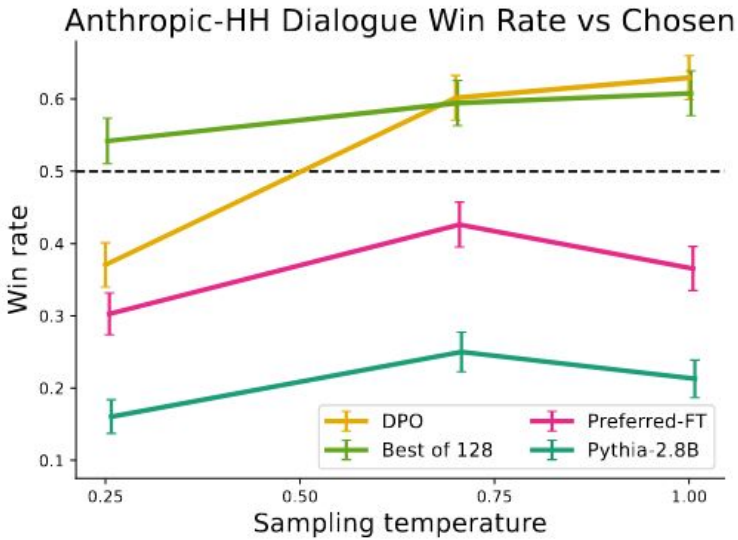
1. Задача: продолжить отзыв на фильм, чтобы он казался позитивным
2. Оцениваем среднюю награду при фиксированном отклонении новой от исходных весов

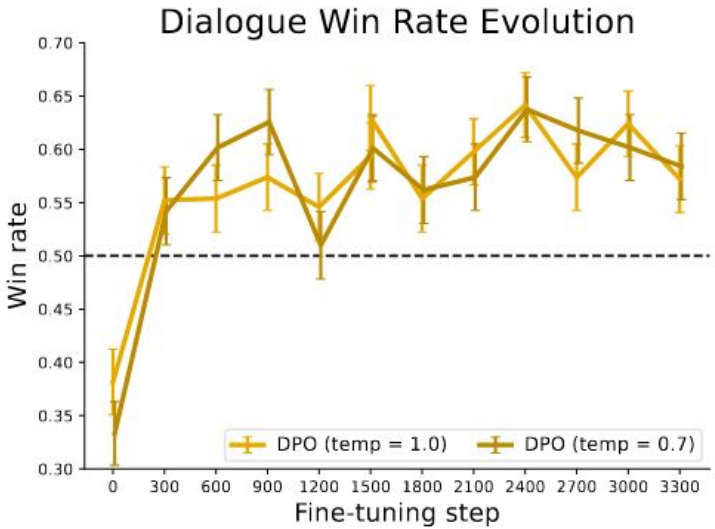


DPO побеждает!

DPO: эксперименты

Задача: ответить на запрос пользователя к LLM. Имеем датасет с (x, y_w, y_l). Win Rate: сравниваем ответ модели и y_w (лучшее выбирает сторонний наблюдатель)





Chain of Hindsight: другая идея

Higher School of Economics
Faculty of Computer science

А может можно еще проще?

Отказываемся от специального процесса дообучения.

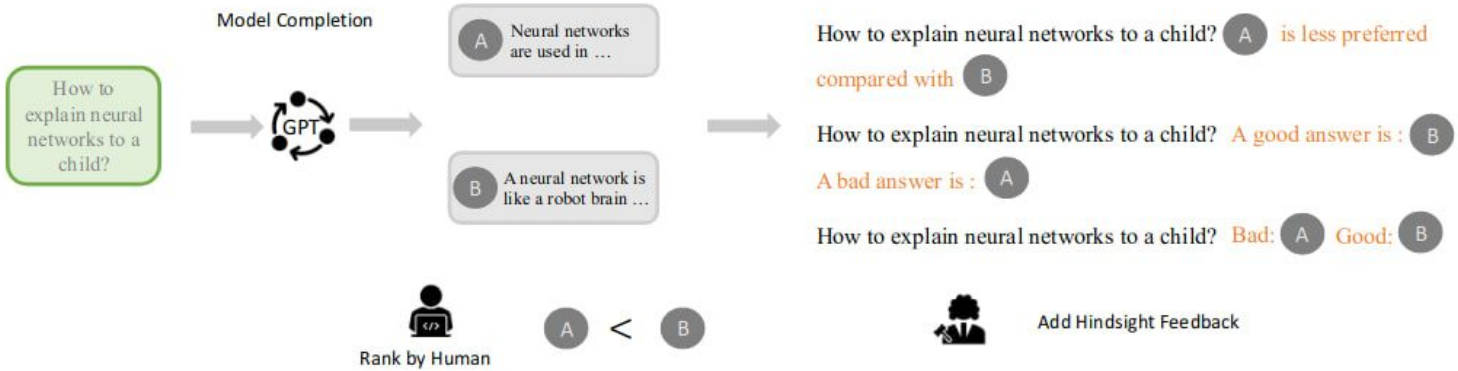
1. Возьмем предобученную модель (аналогично SFT модели раньше)
2. Сгенерируем (или еще как-либо получим) на каждый промпт по два ответа
3. Людьми разметим из каждой пары ответ лучше и ответ хуже
4. Дообучим модель на этих примерах (подробнее далее)

Chain of Hindsight: другая идея

Higher School of Economics
Faculty of Computer science

Имеем выборку (x, y_w, y_l). Как будем дообучать?

Показываем в одном из следующих форматов:



Chain of Hindsight: некоторые детали

Higher School of Economics
Faculty of Computer science

1. Полезно показывать модели сравнения в нескольких форматах
   a. Prompt: … Good: …, Bad…
   b. Prompt: … You’re a helpful assistant: … You’re an unhelpful assistant: …
   c. e.t.c.
2. При обучении маскируем часть токенов в Good response (5-10%). Иначе модель будет просто копировать в Bad часть всю Good часть
3. В Inference важно после промпта писать Good или You’re a helpful assistant, чтобы модель генерировала ответы именно на основе хороших примеров

Chain of Hindsight: эксперименты

Higher School of Economics
Faculty of Computer science

Даны: пары диалогов между человеком и LLM, где людьми в каждой паре выбраны более удачные. **Задача:** классификация, какой диалог удачнее.

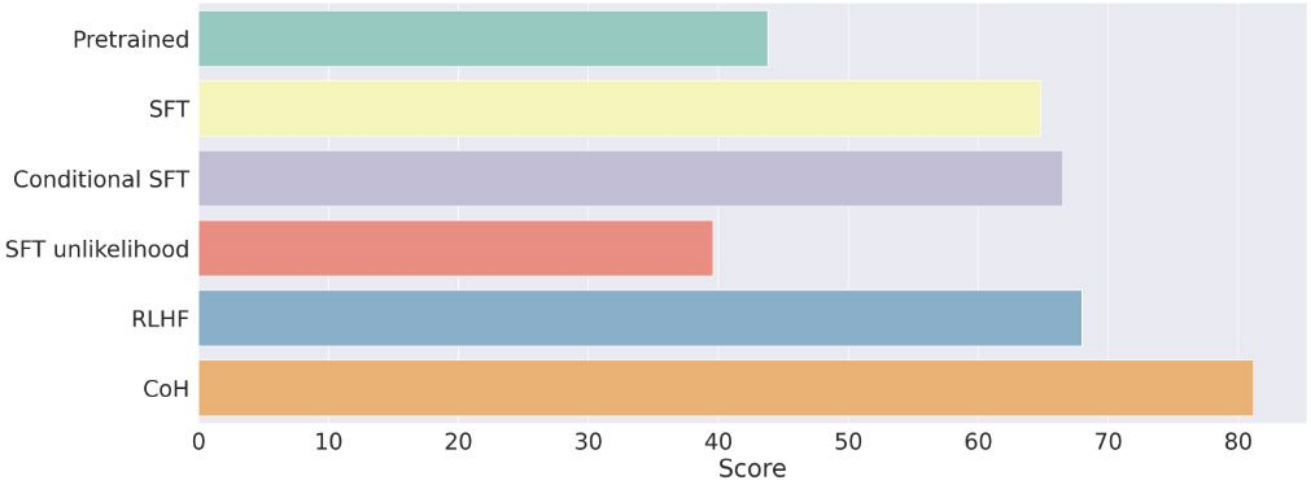


Figure 4: **Evaluation on dialogue**. Comparing CoH with RLHF and SFT baselines. The metric is the accuracy of classifying the preferred dialogue.

Chain of Hindsight: еще эксперименты

Higher School of Economics
Faculty of Computer science

Задача: по тексту генерировать краткое содержание (summary). **Тексты:** отфильтрованные посты из Reddit. **Метрика:** оценки людей по критериям.

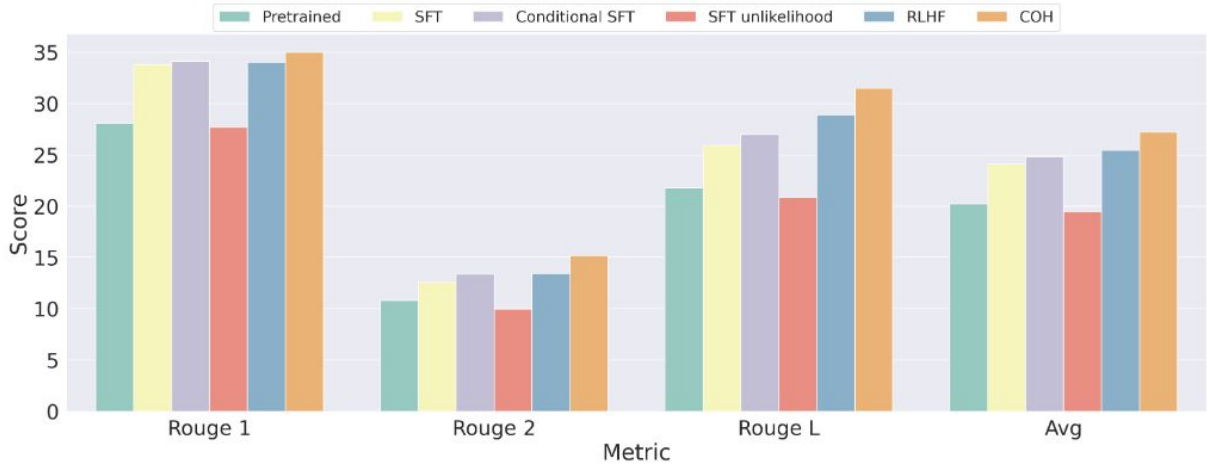


Figure 3: **Evaluation on summarization**. Comparison between RLHF, SFT and CoH. The metrics are ROUGE scores on TL;DR summarization task.

Заключение

Higher School of Economics
Faculty of Computer science

На практике часто все еще применяется RLHF, но новые методы пытаются от него отдалиться из-за сложности и стоимости обучения в подходе RLHF.

DPO и CoH: наиболее удачные попытки отойти от RLHF.

DPO: с помощью математики избавляемся от необходимости в reward model

CoH: работаем над промптами, которые показываем предобученной LLM

Важно: это все про fine-tuning!

Ссылки (где можно почитать подробнее)

Higher School of Economics
Faculty of Computer science

- RLHF: https://doi.org/10.48550/arXiv.1909.08593
- DPO: https://doi.org/10.48550/arXiv.2305.18290
- CoH: https://arxiv.org/abs/2302.02676
- Про RLHF в HuggingFace: https://huggingface.co/docs/trl/

Спасибо за внимание!