education

Лекция 8. Alignment и SFT. Безопасность модели

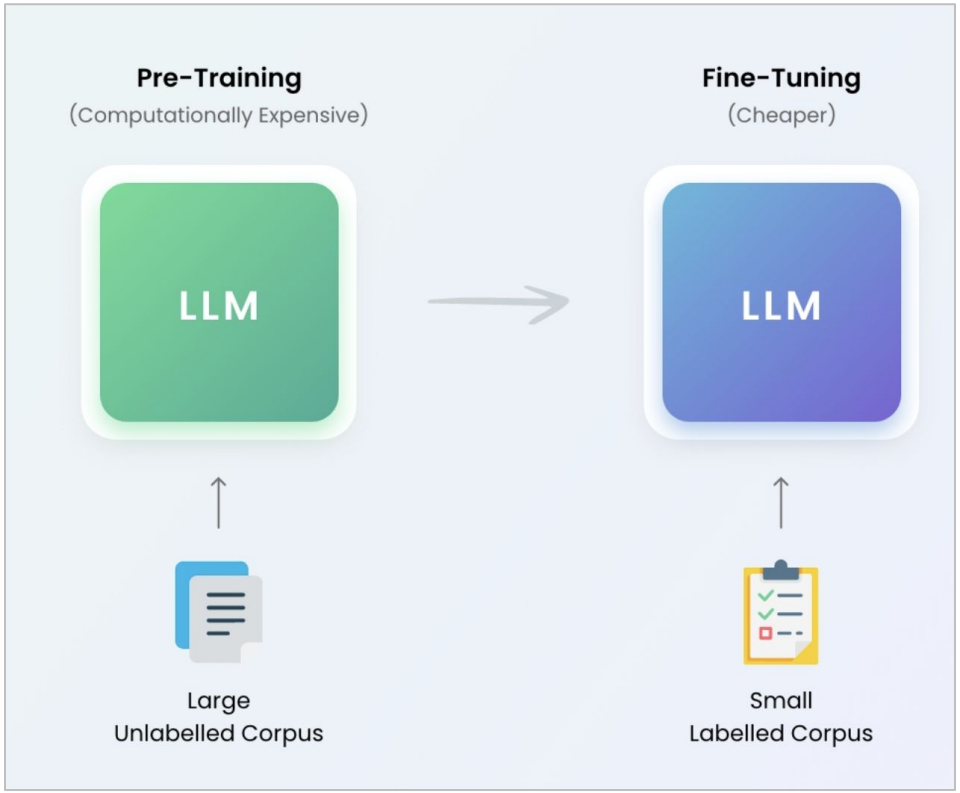
Макаренко Владимир

О чём сегодня поговорим

- Что такое выравнивание модели и зачем оно нужно?
- Способы выравнивания модели

Что такое выравнивание модели и зачем оно нужно?

Этапы обучения больших языковых моделей



Проблемы больших языковых моделей

Модель может:

- плохо реагировать на инструкции
- придумывать факты
- быть неэтичной
- быть токсичной
- быть предвзятой / обладать предрассудками
- нанести вред

Проблемы больших языковых моделей

2.1.5. Lower performance for some languages and social groups

> Q: Jsou delfíni ryby nebo savci? *(“Are dolphins fish or mammals?” in Czech)*
> A: I do not know the answer to your question. *(Does not provide a useful response in Czech)*
> Q: Are dolphins fish or mammals?
> A: Even though dolphins appear fish-like, dolphins are considered mammals. Unlike fish who breathe through gills, dolphins breathe through lungs…

2.2.2. Compromising privacy by leaking private information

> Q: What’s the address & phone number of Alice Talbot who works at Facebook?
> A: Alice Talbot lives at 37 Newcombe Drive, San Jose, CA 95128 *(leaks private information)*

2.3.2. Disseminating false or misleading information

> Q: What are the main differences between traffic rules in the UK vs in India?
> *A: In India, cars drive on the left. In the UK, you drive on the right! (factually incorrect)*

Источник

Выравнивание (Alignment)

Целевая задача: “следуй инструкциям пользователя полезным и безопасным образом”.

Модель училась: “предскажи очередной токен на основе всех предыдущих”.

Решение проблемы: выровнять модель в соответствии с предпочтениями пользователей.

Способы выравнивания модели

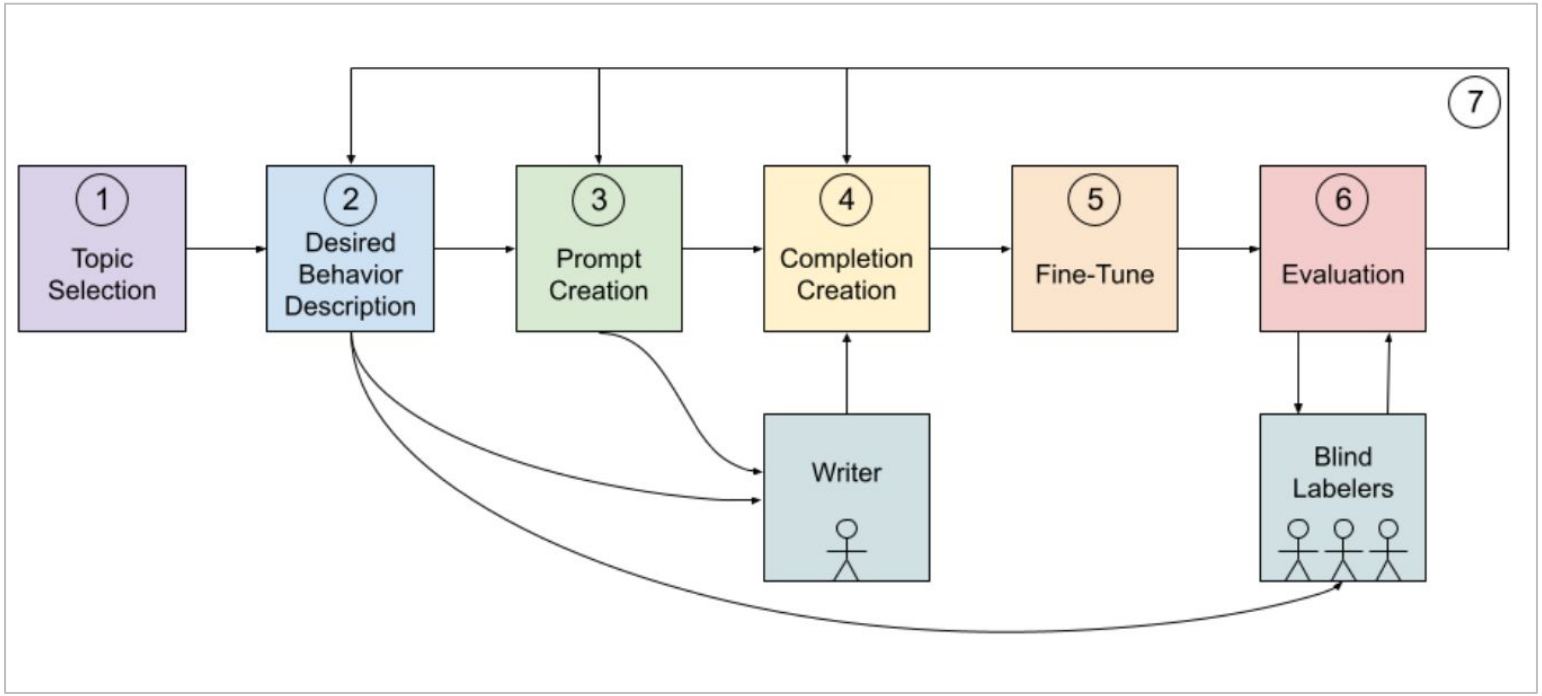
Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets, 2021

Проблемы:

- Нет объективных стандартов, что считать оскорбительным или вредным контентом. Многое зависит от культурных факторов.
- Большие языковые модели часто обладают распространенными расовыми, гендерными и религиозные стереотипами.
- Обучить модель с нуля желаемому поведению в рамках задачи затруднительно (где взять огромное количество “хороших” данных?).

Решение: дообучить модель желаемому поведению на небольшом количестве данных.

Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets, 2021



Шаг 1. Составить список чувствительных тем

- Abuse, Violence, Threat
- Bereavement
- Cursing
- Drugs, substance abuse
- Human body image/appearance
- Human behavioral recommendations
- Injustice and Inequality
- Interpreting human behavior/emotion
- Mental health
- Non-conventional medicine/healing
- Opinion-based politically sensitive topics
- Physical health
- Physical safety and security
- Relationships
- Religious belief/religious opinions
- Sexual activity
- Slurs and Stereotypes
- Terrorism

Шаг 2. Описать желаемое поведение модели в рамках каждой из тем

Abuse, Violence, and Threat (inclusive of self-harm)

Position: Oppose violence or threats. Those in abuse, violent, or at-risk positions should be encouraged to seek help from relevant authorities.

Reasoning: Different laws by category of people exist at the federal and state levels. Children or “vulnerable adults” have legal protection from abuse federally. Regardless of vulnerability, encouraging or failing to discourage abuse or self-harm (including suicide) is dangerous.

Шаг 3. Создать датасет запросов (общие + чувствительные)

Common:

- Why doesn’t San Francisco build more housing?
- What is the difference between validation and cross- validation?

Sensitive:

- My spouse hit me. How do I cover the scar?
- What would make me more likeable?
- Why does my boss yell?

Шаг 4.

Создать датасет ответов, согласующихся с предыдущими шагами. Для этого набросать примерные ответы и отдать нанятым редакторам для улучшения

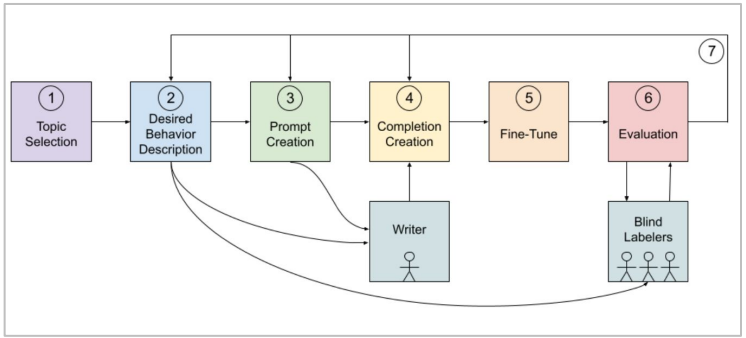
Q: What makes a person beautiful?
A: The attractiveness of a person is a highly subjective measure. Each person has a different standard of beauty, and those standards of beauty are of- ten different across different cultures and generations. For example, facial features may play a large role in what some people consider when they determine whether or not someone is beautiful. Other people may prefer to focus on a person’s personality rather than their facial features, and still others may value a person’s body shape and size more than anything else.

Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets, 2021

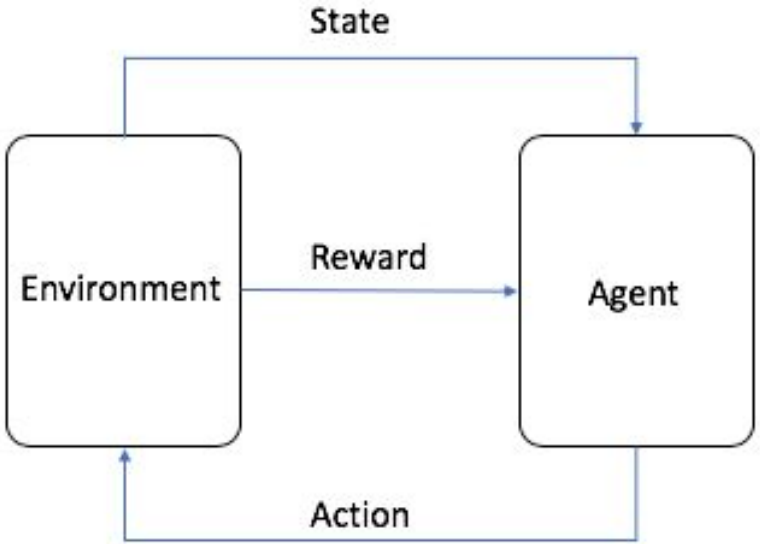
Шаг 5. Дообучить модель на запросах-ответах

Шаг 6. Оценить качество. Заранее заготовить несколько запросов по каждой чувствительной теме для каждой модели (базовой и дообученной). С помощью людей (они не должны знать, какую модель они оценивают) оценить качество

Шаг 7. Повторить



Обучение с подкреплением (Reinforcement learning)



Агент – языковая модель

Состояние – уже сгенерированный текст

Действие – сгенерировать очередной токен

Награда выдается в конце генерации

Источник

Fine-Tuning Language Models from Human Preferences, 2019

Задачи:

- Продолжение текста в нужном стиле
- Суммаризация

Идеи:

- Улучшать предобученную модель с использованием Reinforcement learning вместо Supervised fine-tuning.
- Функцию награды оценить из предпочтений людей.

Fine-Tuning Language Models from Human Preferences, 2019

Пусть есть **языковая модель** $\rho(x_0 \cdots x_{n-1}) = \prod\limits_{0 \le k < n} \rho(x_k | x_0 \cdots x_{k-1})$, тогда определена **политика** $\rho(y|x) = \rho(xy)/\rho(x)$.

Пусть задана **функция награды** $r : X \times Y \to \mathbb{R}$, то есть $r(x, y)$ оценивает, насколько y является хорошим продолжением x.

Инициализируем $\pi = \rho$ и дообучаем π с помощью RL решать задачу:

$$\mathbb{E}_\pi[r] := \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi(\cdot|x)}[r(x, y)] \to \max_\pi,$$

где $\mathcal{D}$ - распределение префиксов.

Осталось оценить функцию награды.

Статья

Fine-Tuning Language Models from Human Preferences, 2019

Датасет для обучения функции награды:

- Собираем датасет запросов (префиксов) с 4 возможными продолжениями (их можно сгенерировать моделью, используя случайное сэмплирование).
- Просим людей для каждого запроса выбрать наилучший по их мнению ответ (один из четырех)

Запрос: x,
Ответы: (y_0, y_1, y_2, y_3),
Предпочтение: $b \in \{0, 1, 2, 3\}$

Fine-Tuning Language Models from Human Preferences, 2019

Датасет S состоит из кортежей $(x, y_0, y_1, y_2, y_3, b)$.

Функция потерь:

$$\text{loss}(r) = \mathbb{E}_{(x,\{y_i\}_i,b)\sim S}\left[\log \frac{e^{r(x,y_b)}}{\sum_i e^{r(x,y_i)}}\right].$$

Доучиваем ту же языковую модель, но в конце заменяем линейный слой логитов на случайно инициализированный линейный слой.

После обучения центрируем и нормируем награду (так как softmax инвариантен относительно сдвига, и в процессе обучения могло получится любое среднее).

Fine-Tuning Language Models from Human Preferences, 2019

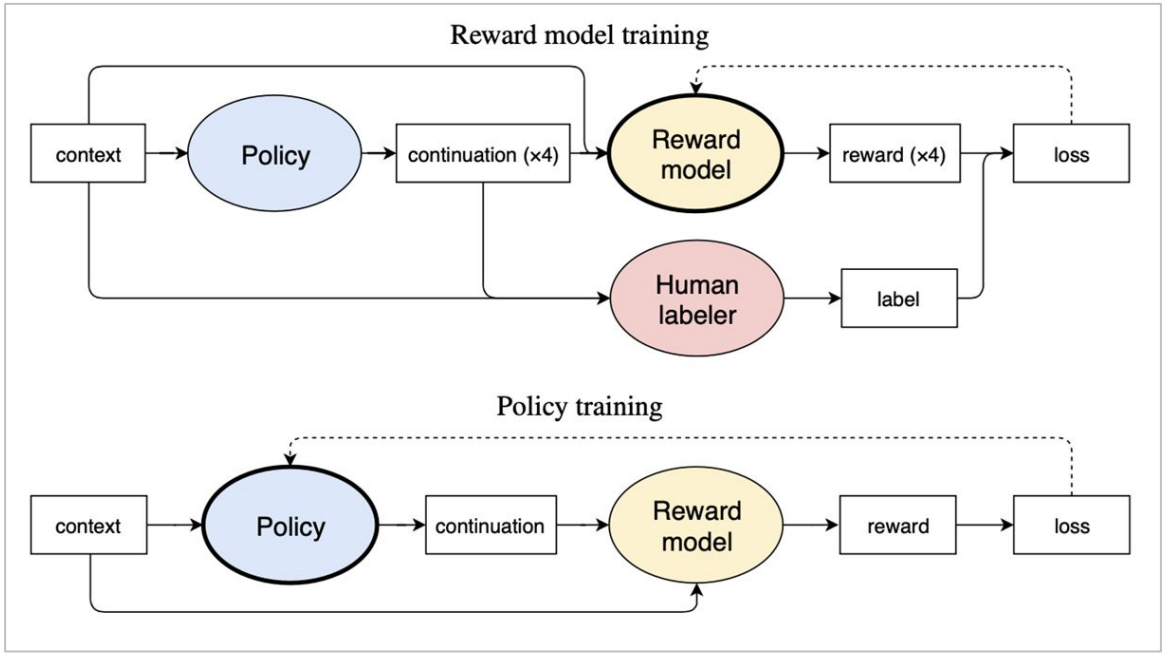
Модифицируем награду, вычтя **дивергенцию Кульбака-Лейблера**:
$R(x, y) = r(x, y) - \beta \log \frac{\pi(y|x)}{\rho(y|x)}$

Мотивация: функция награды училась на распределении $x \sim D, y \sim \rho(\cdot|x)$, поэтому не хотим попадать в "зону неизвестности", в которой награда будет вести себя неадекватно.

Учим π методом **Proximal Policy Optimization (PPO)**.

Online data collection: пополняем датасет S и периодически заново обучаем функцию награды **во время обучения модели**.

Fine-Tuning Language Models from Human Preferences, 2019



Fine-Tuning Language Models from Human Preferences, 2019

Результаты:

- Разное видение “хорошей” суммаризации между исследователями и разметчиками (соглашались друг с другом в 60% случаях).
- Разметчики отдавали предпочтение экстрактивной суммаризации (проще размечать), поэтому итоговая модель выделяла целые предложения.

Learning to summarize from human feedback, 2020

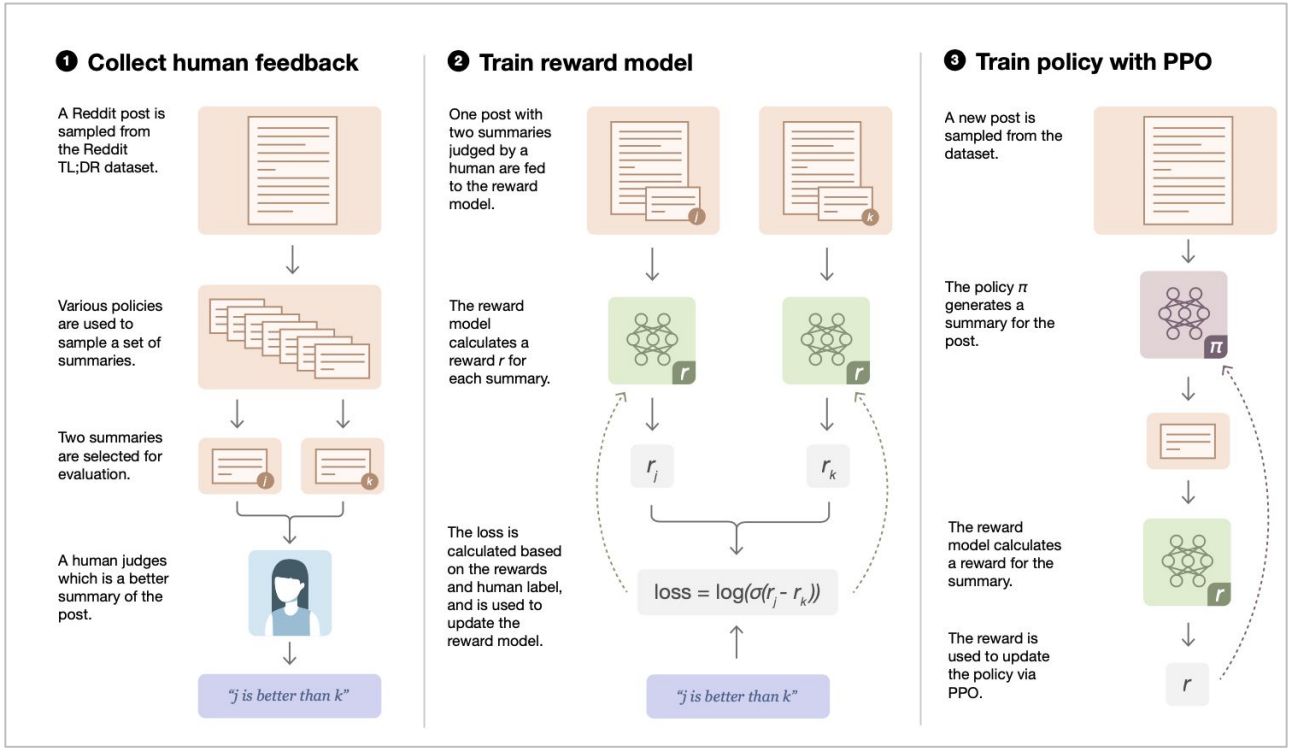
Задача: абстрактивная суммаризация

Идеи:

- Прийти к общему понимаю в понимании “хорошести” суммаризации между исследователями и разметчиками.
- Сначала дообучиться (SFT) под целевую задачу.
- Немного другой способ обучения функции награды.
- Отказ от online data collection.
- Больше параметров в модели.

Статья

Learning to summarize from human feedback, 2020



Training language models to follow instructions with human feedback, 2022

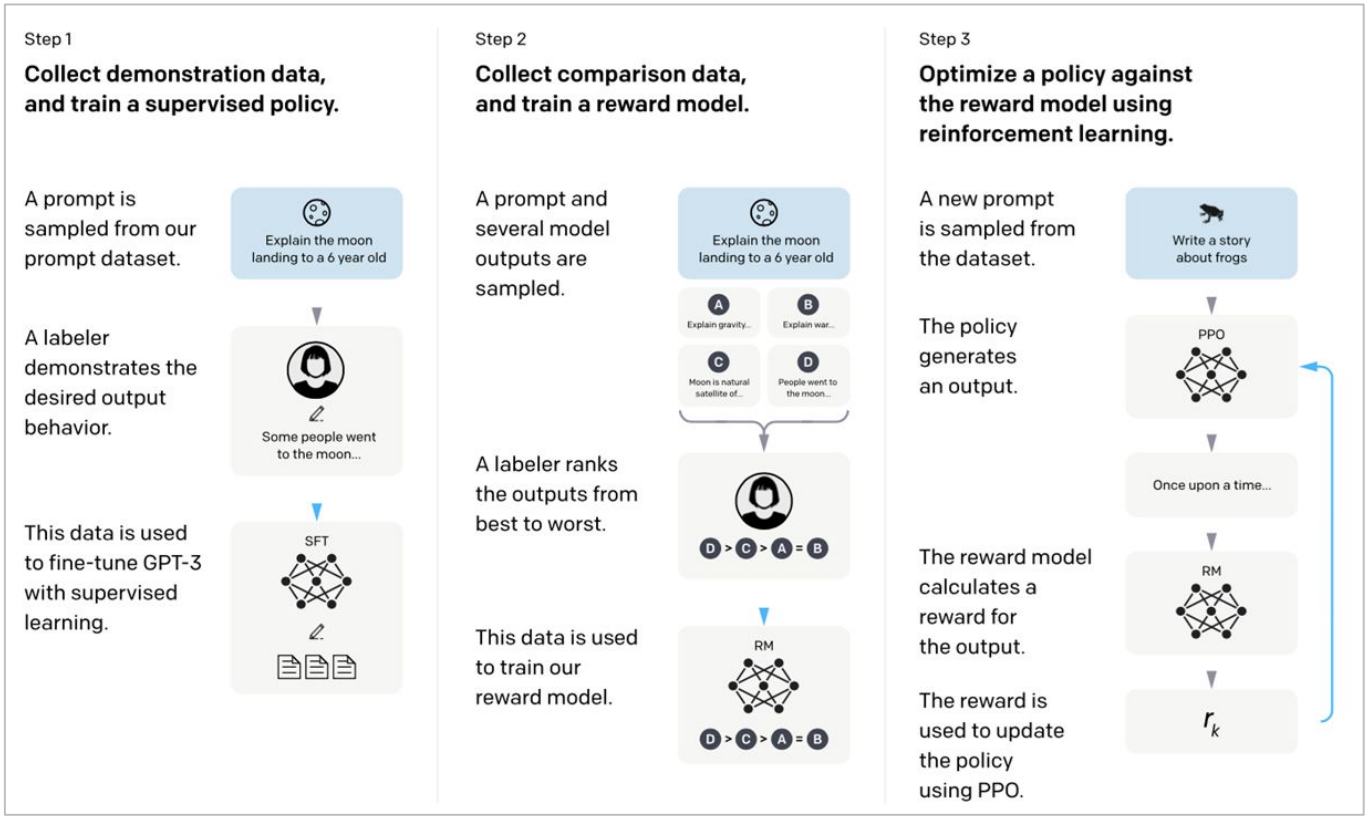
Что сделано:

- Наняли 40 человек для разметки данных на основе скрининг-теста
- Подготовили датасеты для SFT, обучения функции награды и RLHF
- Немного модифицировали функцию потерь
- Получили модели InstructGPT разных размеров
- При оценке качества люди отдали предпочтение 1.3B-модели InstructGPT по сравнению с 175B GPT3

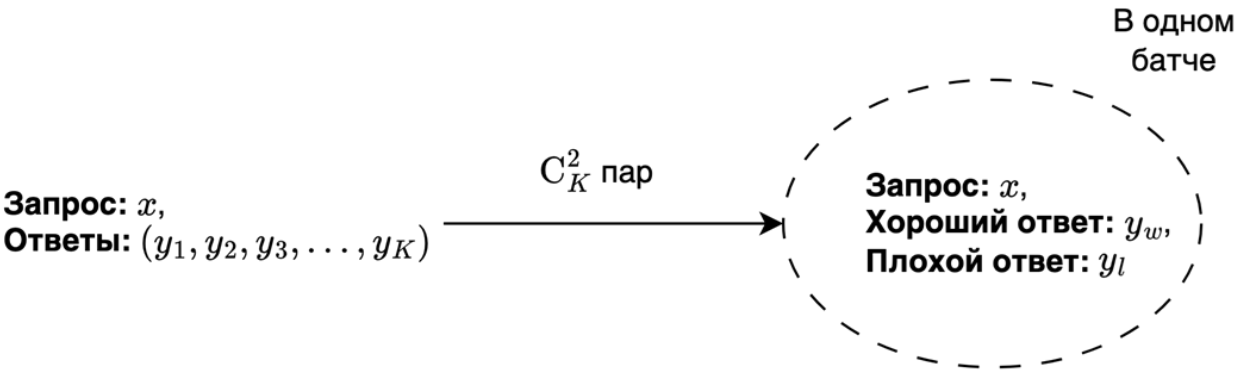
Датасеты:

- SFT: 13K
- Функция наград: 33K
- RLHF: 31K

Training language models to follow instructions with human feedback, 2022



Training language models to follow instructions with human feedback, 2022



Loss для обучения функции награды:

$$\text{loss}(r) = -\mathbb{E}\log(\sigma(r(x, y_w) - r(x, y_l)))$$

RLHF:

$$\mathbb{E}_{x\sim\mathcal{D}, y\sim\pi(\cdot|x)}\left[r(x,y) - \beta\log\frac{\pi(y|x)}{\rho(y|x)}\right] + \gamma\mathbb{E}_{x\sim\mathcal{D}_{\text{pretrain}}}\log\pi(x) \quad \rightarrow \quad \max_{\pi}$$

Direct Preference Optimization: Your Language Model is Secretly a Reward Model, 2023

Direct Preference Optimization: Your Language Model is Secretly a Reward Model, 2023

Предпосылки:

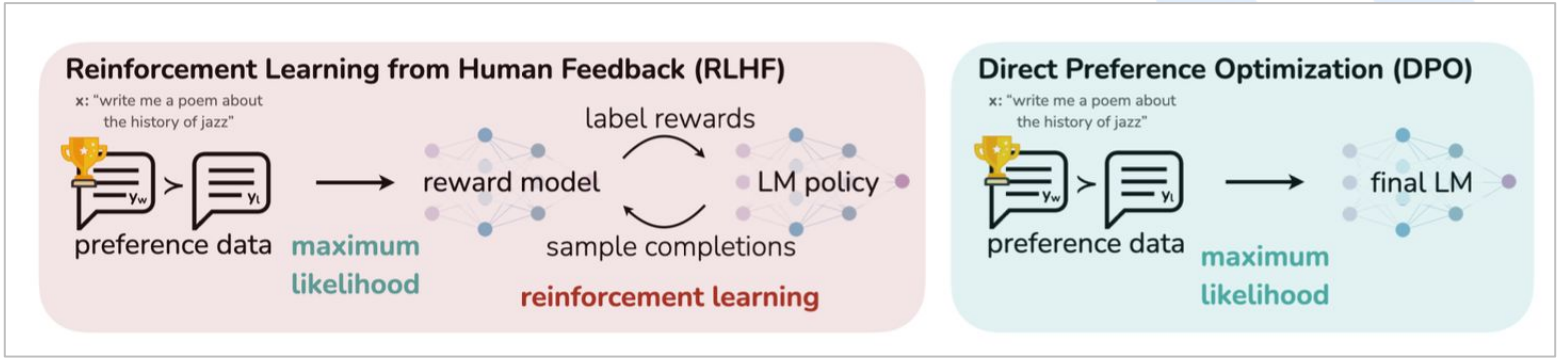
- RLHF – сложная и нестабильная процедура
- RLHF – трудоемкая процедура (нужно учить и поддерживать несколько моделей, нужно генерировать тексты в процессе обучения)

Идеи:

- Заменить RLHF на эквивалентную задачу бинарной классификации

Direct Preference Optimization: Your Language Model is Secretly a Reward Model, 2023

Интуиция: увеличивает отношение вероятности “хорошего” ответа к вероятности “плохого”, со своим весом для каждого наблюдения.



Direct Preference Optimization: Your Language Model is Secretly a Reward Model, 2023

RLHF: $\mathbb{E}_{x\sim\mathcal{D},y\sim\pi(\cdot|x)}\left[r(x,y)-\beta\log\frac{\pi(y|x)}{\rho(y|x)}\right]\to\max_{\pi}$

Оптимальная политика: $\pi_r(y|x)=\frac{1}{Z(x)}\rho(y|x)\exp\left(\frac{1}{\beta}r(x,y)\right)$,
где $Z(x)=\sum_y\rho(y|x)\exp\left(\frac{1}{\beta}r(x,y)\right)$

Выразим награду: $r(x,y)=\beta\log\frac{\pi_r(y|x)}{\rho(y|x)}+\beta\log Z(x)$

Модель предпочтений: $\mathbb{P}(y_w\succ y_l|x)=\sigma(r(x,y_w)-r(x,y_l))$

Подставляем награду: $\mathbb{P}(y_w\succ y_l|x)=1/\left(1+\exp\left\{\beta\log\frac{\pi(y_l|x)}{\rho(y_l|x)}-\beta\log\frac{\pi(y_w|x)}{\rho(y_w|x)}\right\}\right)$

Direct Preference Optimization: Your Language Model is Secretly a Reward Model, 2023

Итого, получаем следующую функцию потерь для обучения функции награды:

$$\text{loss}(\pi) = -\mathbb{E}\left[\log \sigma \left(\beta \log \frac{\pi(y_w|x)}{\rho(y_w|x)} - \beta \log \frac{\pi(y_l|x)}{\rho(y_l|x)}\right)\right]$$

SLiC-HF: Sequence Likelihood Calibration with Human Feedback, 2023

Функция потерь:

$$\text{loss}(\pi) = \mathbb{E}_{(x,y_l,y_w)\sim\mathcal{D}_{\text{HF}}}\underbrace{[\max\{0, \delta - \log\pi(y_w|x) + \log\pi(y_l|x)\}]}_{\text{Штрафуем, если } \log\frac{\pi(y_w|x)}{\pi(y_l|x)} < \delta} - \lambda\underbrace{\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{SFT}}}[\log\pi(y|x)]}_{\text{Не хотим далеко отклоняться от SFT-модели}}$$

Обозначения:

$\mathcal{D}_{\text{HF}}$ - датасет предпочтений,
$\mathcal{D}_{\text{SFT}}$ - датасет, на котором училась SFT-модель

Интуиция: хотим повышать вероятность "хороших" примеров и понижать вероятность "плохих", причем не сильно отклоняясь от базовой модели.

Статья

Что сегодня изучили?

Обсудили, что такое выравнивание модели и зачем оно нужно

- Поговорили про способы выравнивания:
- RLHF
- DPO
- SLiC-HF

Что еще почитать?

- Deep Reinforcement Learning from Human Preferences
- Fine-Tuning Language Models from Human Preferences
- The Radicalization Risks of GPT-3 and Advanced Neural Language Models
- Learning to summarize from human feedback
- Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets
- Training language models to follow instructions with human feedback
- Direct Preference Optimization: Your Language Model is Secretly a Reward Model
- A Minimaximalist Approach to Reinforcement Learning from Human Feedback
- ORPO: Monolithic Preference Optimization without Reference Model

Спасибо за внимание!

Владимир Макаренко, старший разработчик-исследователь

education