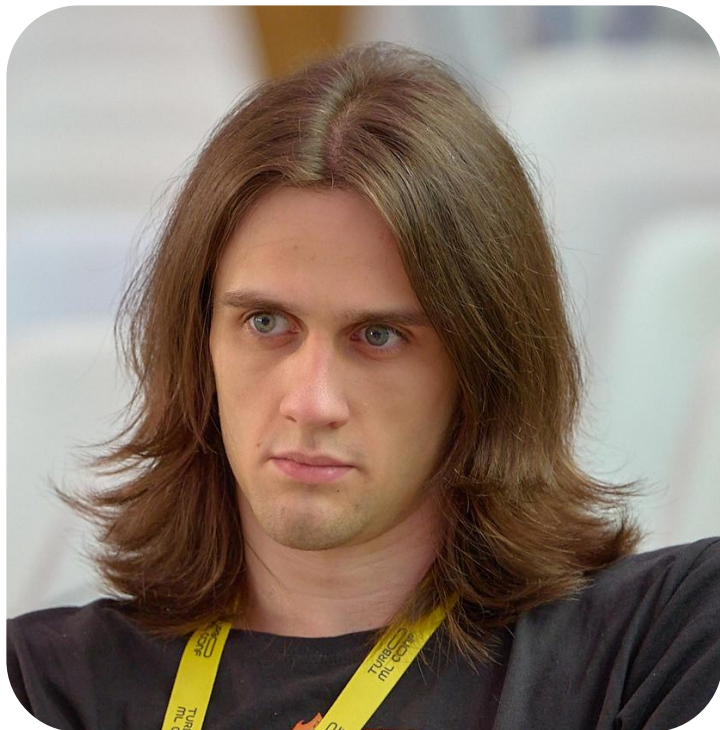


Точка для
предпринимателей
и предприятий

Alignment for all



Илларион Иов

Занимаюсь дообучением LLM
для внутренних сервисов Точки

Аспирант ИТМО, развиваю
применение LLM в инструментах
AutoML

До этого разрабатывал решения
для суррогатной оптимизации
в Huawei, Bosch

О чём поговорим

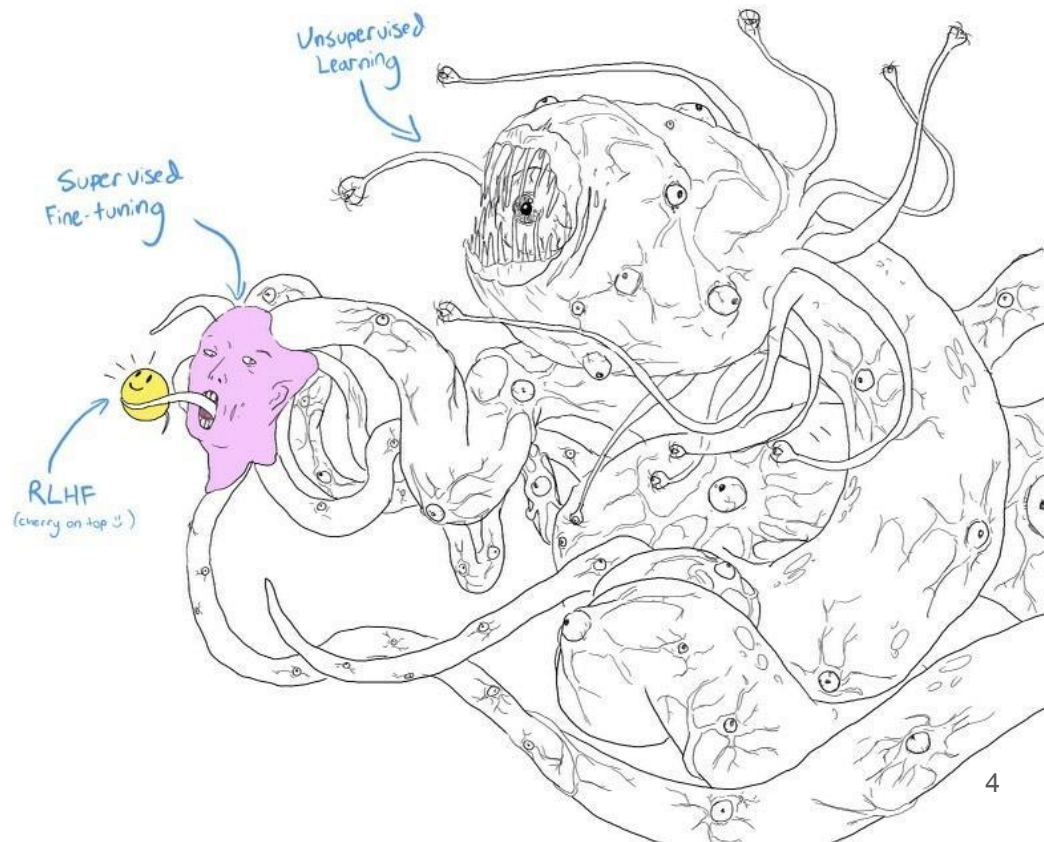
- ✓ Когда нам важны предпочтения
- ✓ Оригинальные методы оптимизации предпочтений
- ✓ Упрощённые методы оптимизации предпочтений
- ✓ Наше решение

Этапы обучения LLM

Уровень модели

Unsupervised Learning
(Pretrain):

Supervised Fine-tuning
(SFT):



**Как понять,
что вам это
нужно?**

Что может пойти не так

- Хорошие показатели на бенчмарках, но в инференсе всё плохо

Что может пойти не так

- Хорошие показатели на бенчмарках, но в инференсе всё плохо
- Модель классификации, обученная для обнаружения токсичных и опасных ответов, часто срывает

Что может пойти не так

- Хорошие показатели на бенчмарках, но в инференсе всё плохо
- Модель классификации, обученная для обнаружения токсичных и опасных ответов, часто срывает
- Высокая дисперсия в ответах, низкие значения $\log(\pi(y|x))$

Что может пойти не так

- Хорошие показатели на бенчмарках, но в инференсе всё плохо
- Модель классификации, обученная для обнаружения токсичных и опасных ответов, часто срывает
- Высокая дисперсия в ответах, низкие значения $\log(p_i(y|x))$
- Энкодер модели показывают низкую близость контекста и ответа

Что может пойти не так

- Хорошие показатели на бенчмарках, но в инференсе всё плохо
- Модель классификации, обученная для обнаружения токсичных и опасных ответов, часто срывает
- Высокая дисперсия в ответах, низкие значения $\log(p_i(y|x))$
- Энкодер модели показывает низкую близость контекста и ответа
- Энкодер модели показывает низкую близость контекста и ответа

Что может пойти не так

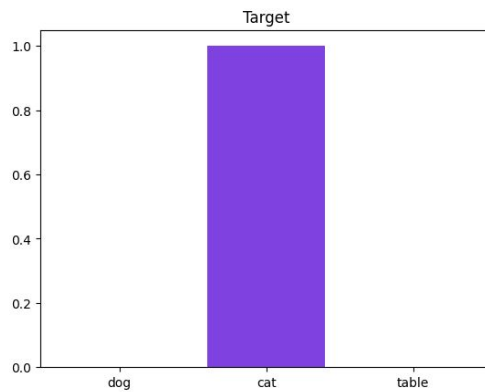
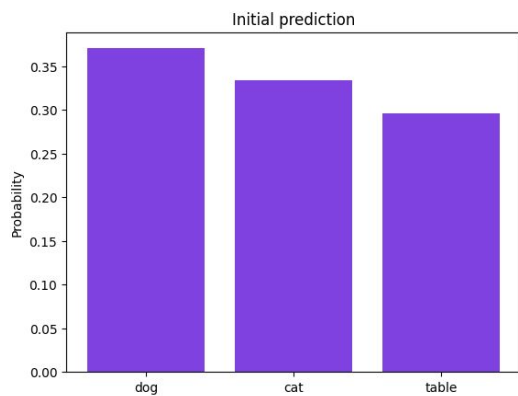
- Хорошие показатели на бенчмарках, но в инференсе всё плохо
- Модель классификации, обученная для обнаружения токсичных и опасных ответов, часто срывает
- Высокая дисперсия в ответах, низкие значения $\log(p_i(y|x))$
- Энкодер модели показывают низкую близость контекста и ответа
- Энкодер модели показывают низкую близость контекста и ответа

SFT недостаточно

Предпочтения

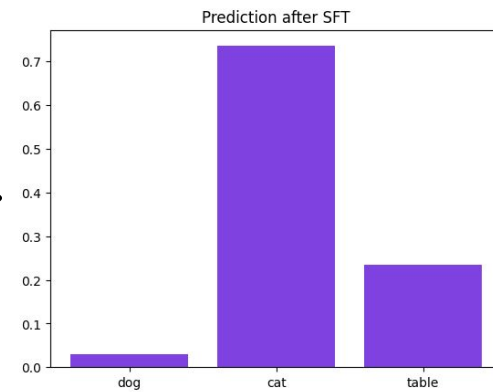
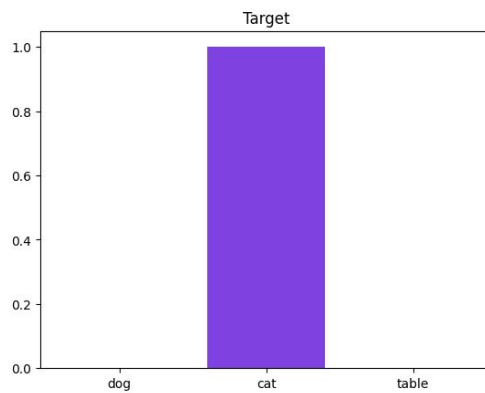
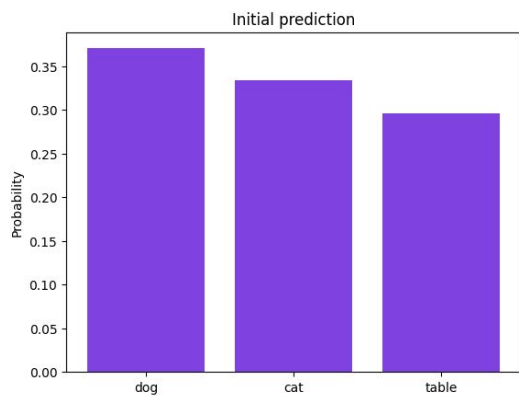
ТОЧКА

LLM: «Я люблю ...»



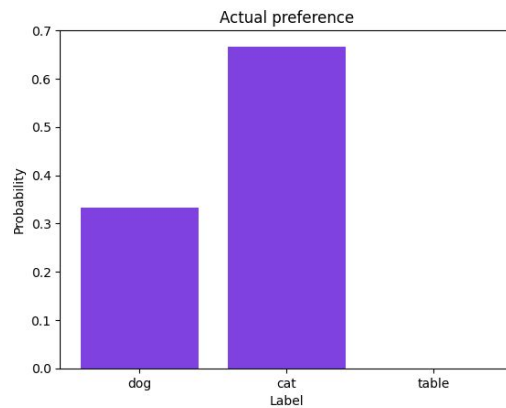
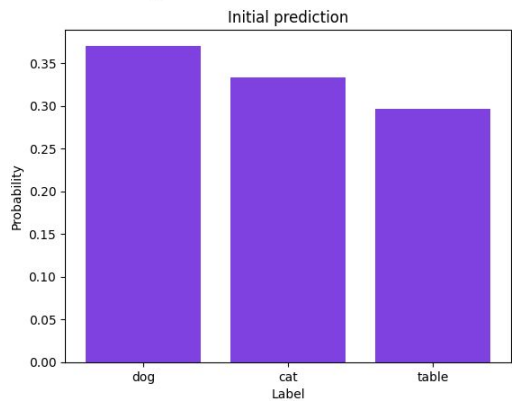
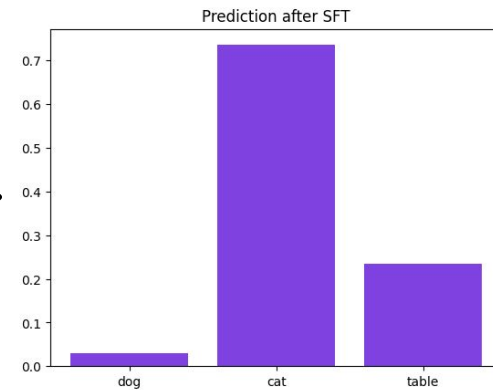
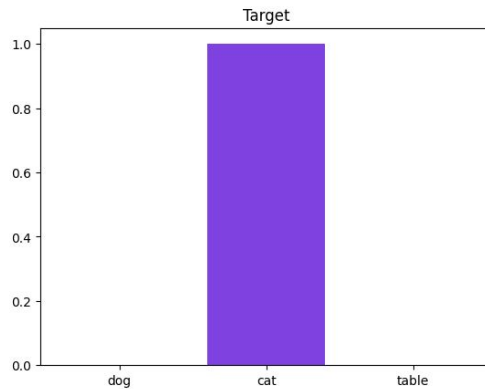
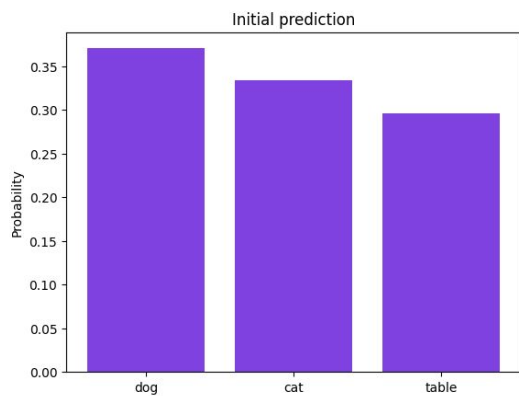
Предпочтения

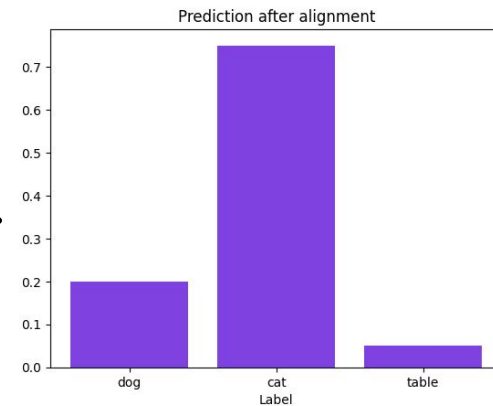
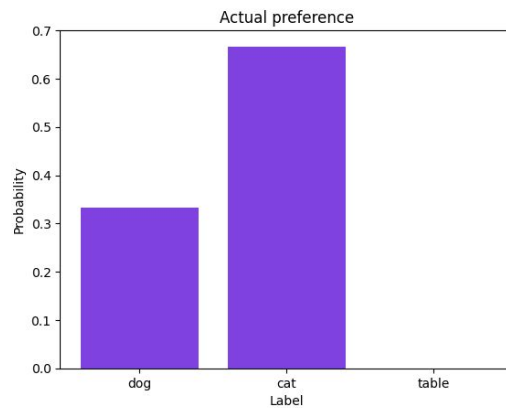
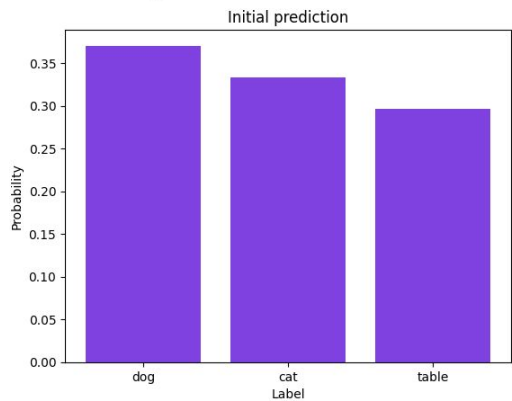
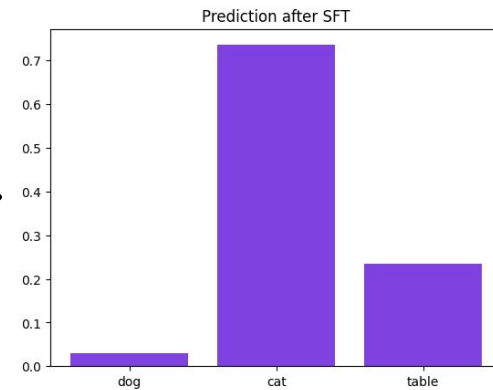
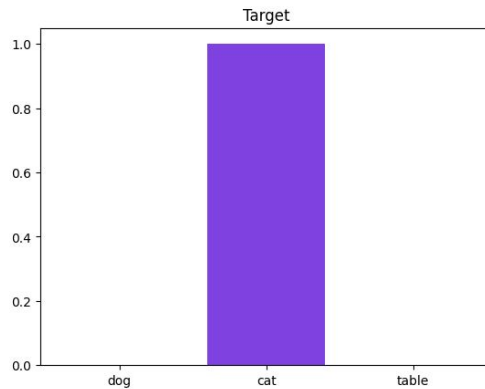
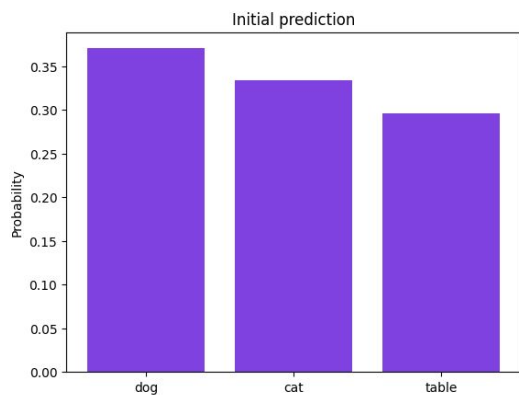
Точка



Предпочтения

ТОЧКА





Этапы обучения LLM

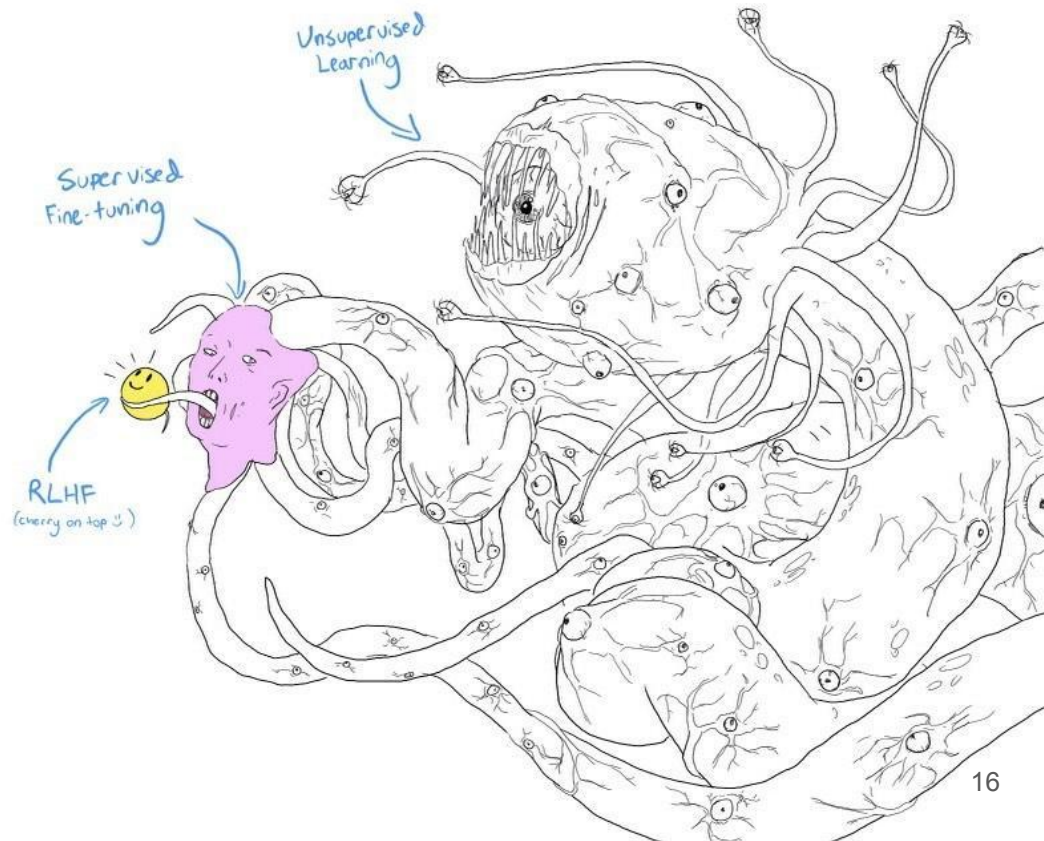
Уровень модели



Unsupervised Learning
(Pretrain):

Supervised Fine-tuning
(SFT):

Preference Optimization
(RLHF):



У вас есть большой GPU кластер?

Да



У вас есть большой штат ассессоров?

Да



**Direct Policy
Optimization
from Human
Feedback**

Direct Reinforcement Learning from Human Feedback

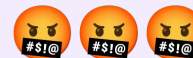
Описание:

- Генерируем ответ на вопрос.
Получаем оценки от ассессоров
- Обучаем модель-агента
получать наибольшую награду

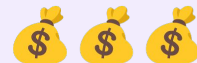
Качество



Сложность



Затраты



Применимость



Direct Reinforcement Learning from Human Feedback

Датасет: набор промптов:

Input 1	Input 2
Опиши структуру атома	Назови 3 главных цвета

Кейсы:

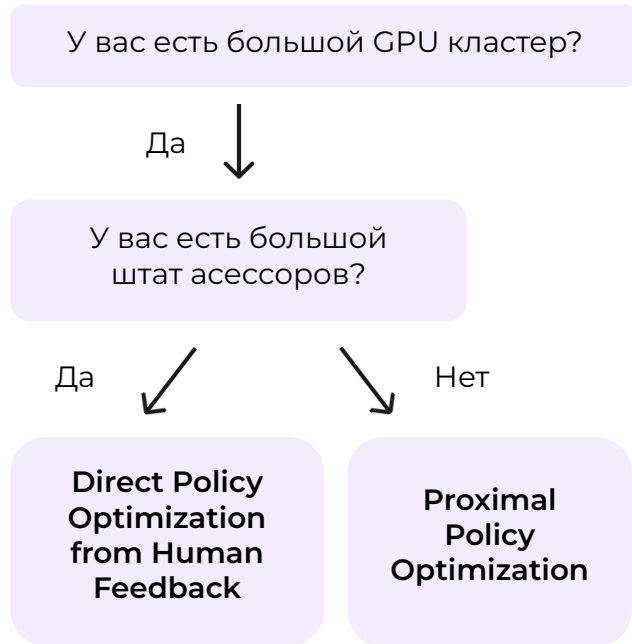
- Маленькие или научные задачи
- Нужно лучшее качество

WHY IS EMPTY???



Выбор метода обучения

ТОЧКА



Proximal Policy Optimization

Описание:

- Сэмплируем ответы из политики
- Обучаем модель награды
- Обучаем модель-агента получать награду выше, чем в исходном датасете

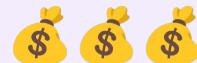
Качество



Сложность



Затраты



Применимость



Proximal Policy Optimization

Датасет: проранжированные ответы

Input	Rank 1	Rank 2	Rank 3
Опиши структуру атома	Атом состоит из ядра...	Ядро, электроны	~_(\ツ)_/~

Кейсы:

- OpenAI GPT-3.5+
- Суммаризация, ответ на вопросы, перевод



Proximal Policy Optimization

Описание:

- Обучаем модель награды
- Обучаем модель-агента получать награду выше, чем в исходном датасете

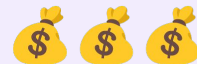
Качество



Сложность



Затраты



Применимость



Proximal Policy Optimization

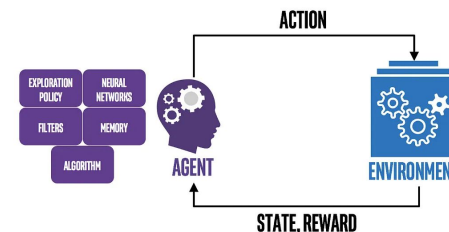
Датасет: оцененные ответы

Input	Rank 1	Rank 2	Rank 3
Опиши структуру атома	Атом состоит из ядра...	Ядро, электроны	~_(\ツ)_/~

Кейсы:

- OpenAI GPT-3.5+
- Суммаризация, ответ на вопросы, перевод

 OpenAI



Начинаем с отношения
вероятностей

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\text{old}}(a_t|s_t)}$$



$$r_t(\theta)$$

Начинаем с отношения
вероятностей

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\text{old}}(a_t|s_t)}$$

Добавим
полученную
«выгоду»

$$r_t(\theta) \cdot \hat{A}_t$$

Начинаем с отношения вероятностей

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\text{old}}(a_t|s_t)}$$

Добавим полученную «выгоду»

$$r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)\hat{A}_t$$

Ограничиваем слишком большие изменения

Начинаем с отношения вероятностей

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\text{old}}(a_t|s_t)}$$

Добавим полученную «выгоду»

$$\min \left(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t \right)$$

В случае малых изменений учитываем их

Ограничиваем слишком большие изменения

Усредняем
по датасету

Начинаем с отношения
вероятностей

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\text{old}}(a_t|s_t)}$$

Добавим
полученную
«выгоду»

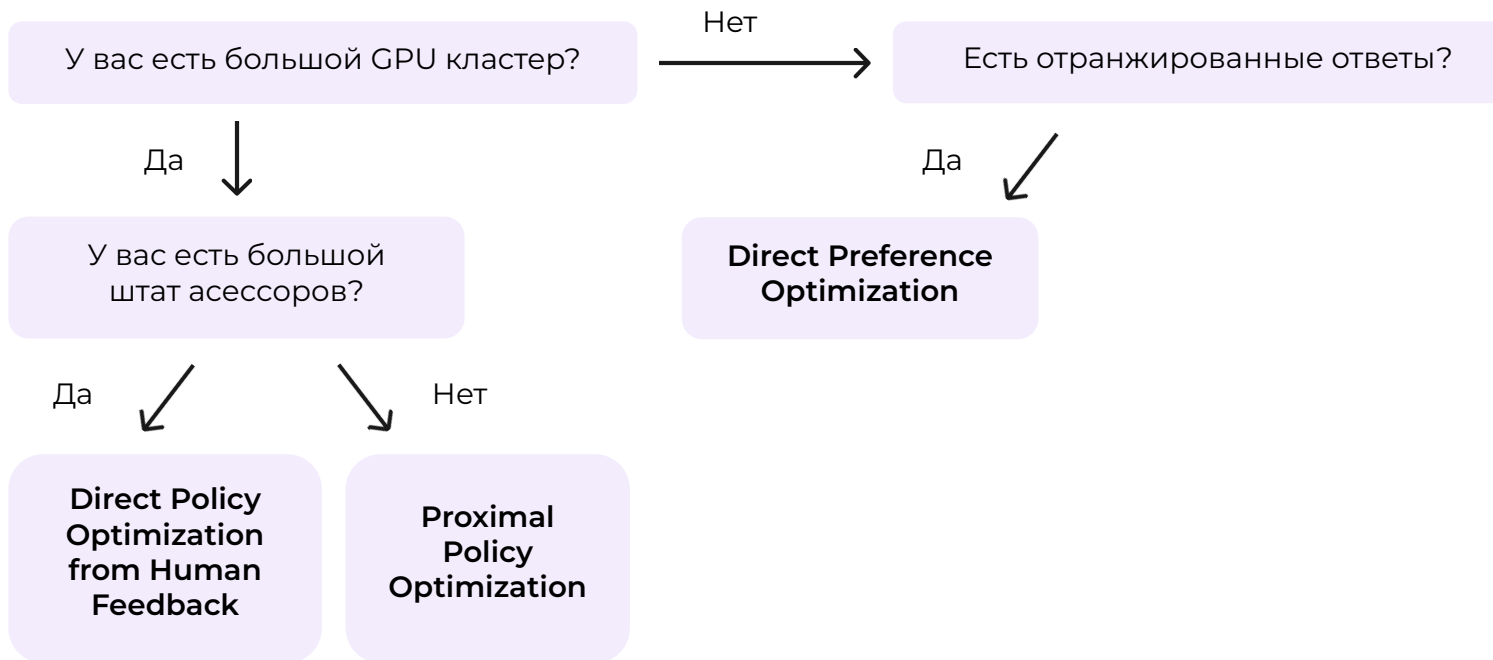
$$\mathcal{L}^{\text{PPO}}(\theta) = \mathbb{E}_t \left[\min \left(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t \right) \right]$$

В случае малых
изменений
учитываем их

Ограничиваем
слишком большие
изменения

Выбор метода обучения

ТОЧКА



Direct Preference Optimization

Описание:

- Обучаем модель генерировать «хороший» ответ с большей вероятностью, чем «плохой»
- Дополнительно ограничиваем сдвиг распределения

Качество



Сложность



Затраты



Применимость



Direct Preference Optimization

Датасет: бинарные предпочтения

Input	Rank 1	Rank 2
Опиши структуру атома	Атом состоит из ядра...	Ядро, электроны

Кейсы:

Открытые модели небольшого размера. Можно дообучать на своих ресурсах

— Llama variations



— Mistral



— Zephyr

— OpenChat

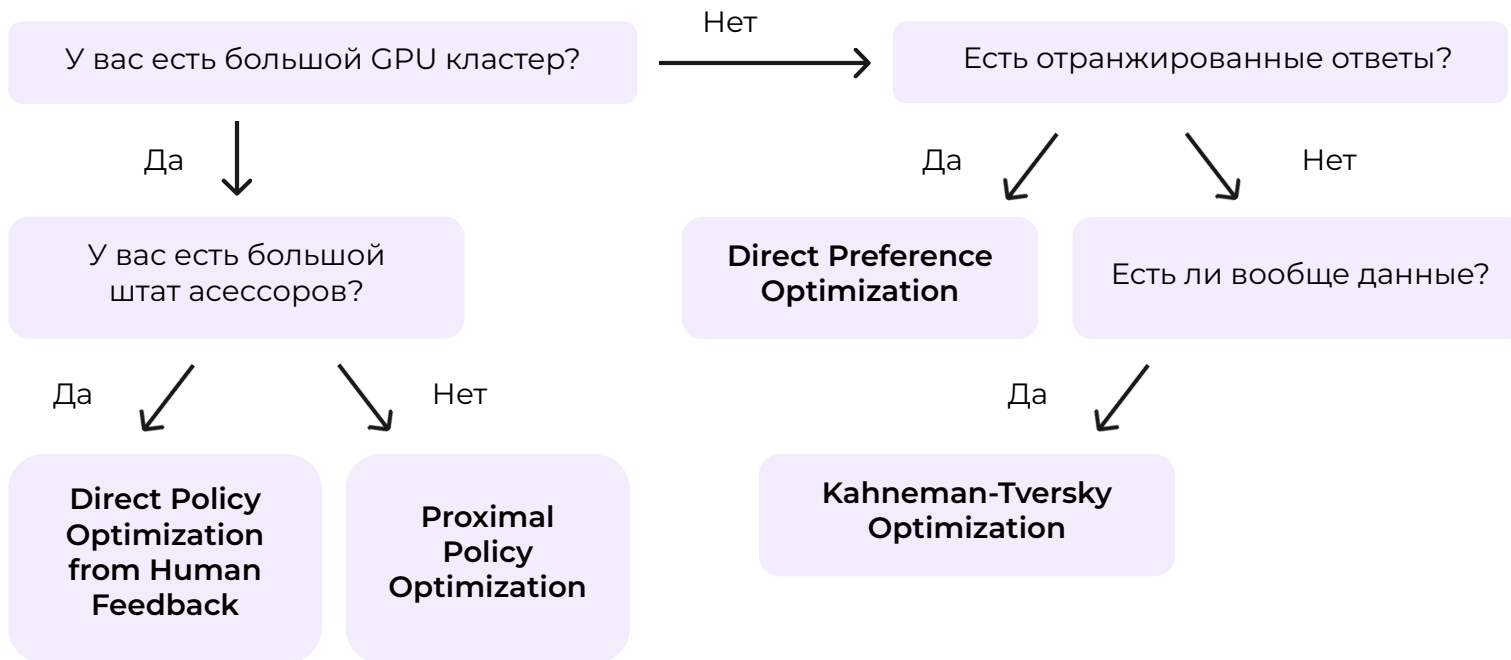
$$\mathcal{L}_{DPO}(\pi_{\theta}; \pi_{ref}) = -\mathbb{E}_{(x; y_w, y_l) \sim D} \left[\log \sigma \left(\beta \log \frac{\pi_{\theta}(y_w|x)}{\pi_{ref}(y_w|x)} - \beta \log \frac{\pi_{\theta}(y_l|x)}{\pi_{ref}(y_l|x)} \right) \right]$$

По сравнению
с начальной моделью
повысилась вероятность
генерации лучшего ответа

По сравнению
с начальной моделью
снизилась вероятность
генерации худших ответов

Выбор метода обучения

ТОЧКА



Kahneman-Tversky Optimization

Вы можете выбирать:

- Получить 1 ₪
- Бросить монетку: орел даёт 5 ₪; решка — 1 ₪

! Игра бесконечная

Kahneman-Tversky Optimization

Вы можете выбирать:

- Получить 1 ₺
- Бросить монетку: орел даёт 5 ₺; решка — 1 ₺

! Единственная попытка

Kahneman-Tversky Optimization

Вы можете выбирать:

- Получить премию
- Бросить монетку: орел даёт премию $\times 5$,
за решку — 1 премия

! Единственная попытка

! Играем на годовую премию

Kahneman-Tversky Optimization

Описание:

Для каждого ответа изменяем вероятность генерации, опираясь на относительную награду от сравнения с исходной моделью

Качество



Сложность



Затраты



Применимость



Kahneman-Tversky Optimization

Датасет: ответы с бинарной характеристикой

Input	Answer	Mark
Опиши структуру атома	Атом состоит из ядра	Ok
Назови 3 главных цвета	Чёрный, белый, синий	Not ok

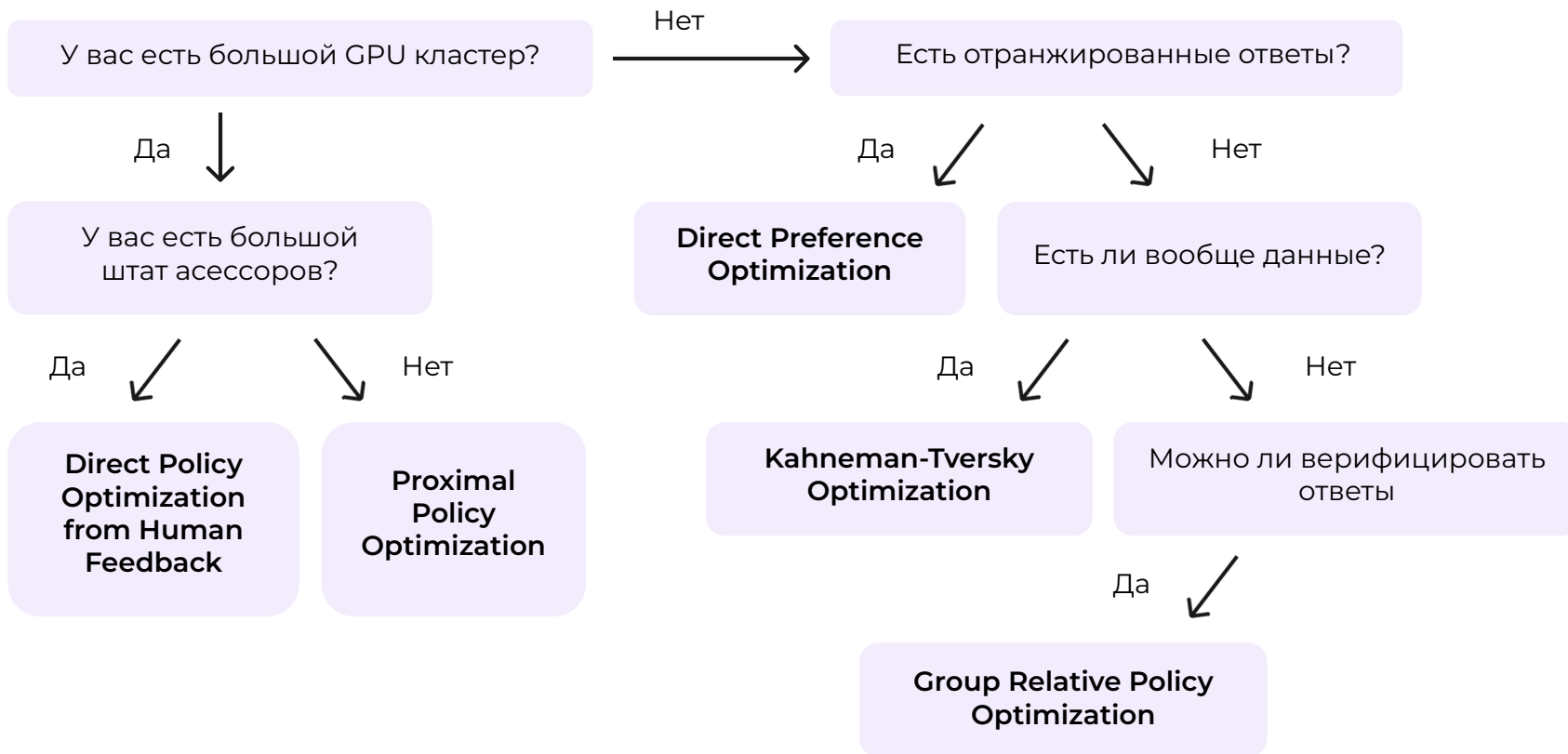
Кейсы:

Наиболее частое применение — безопасность: запрет токсичных ответов, дезинформации, опасных советов



Выбор метода обучения

ТОЧКА



GRPO

Описание:

- Генерируем выборку ответов на вопросы
- Отдельным инструментом находим награду для всех
- Максимизируем среднюю награду по выборке

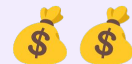
Качество



Сложность



Затраты



Применимость



GRPO

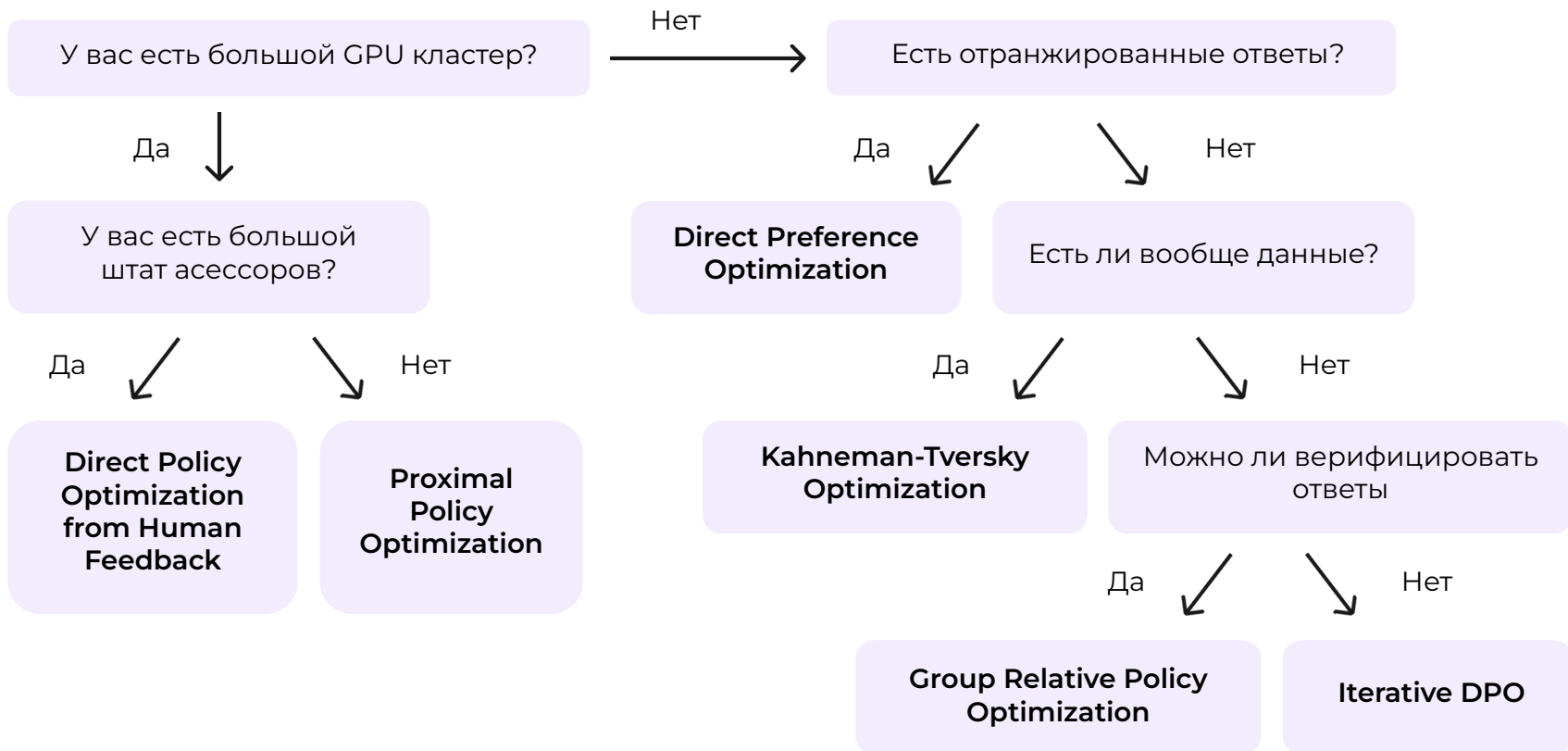
Датасет: набор промптов:

Input 1	Input 2
Опиши структуру атома	Назови 3 главных цвета

Кейсы:

- Логические задачи, рассуждения, код и математика
- Deepseek R1





Iterative Direct Preference Optimization

Описание:

- Генерируем ответы LLM и ранжируем обученной моделью наград
- Обучаем любым методом оптимизации предпочтений
- Повторяем цикл

Качество



Сложность



Затраты



Применимость



Iterative Direct Preference Optimization

Датасет: набор промптов:

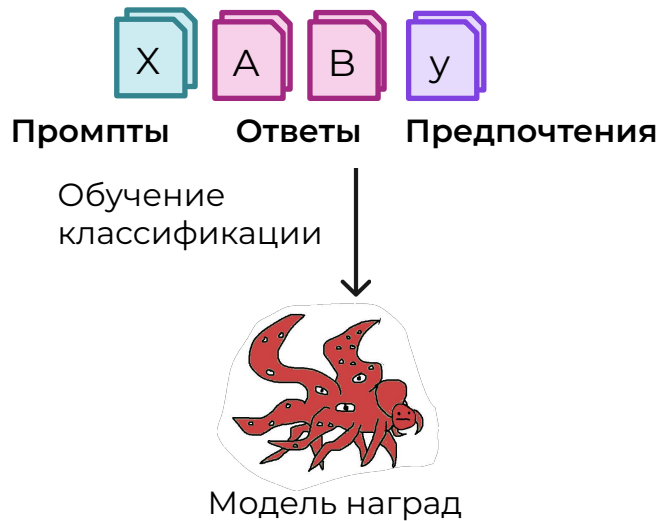
Input 1	Input 2
Опиши структуру атома	Назови 3 главных цвета

Кейсы:

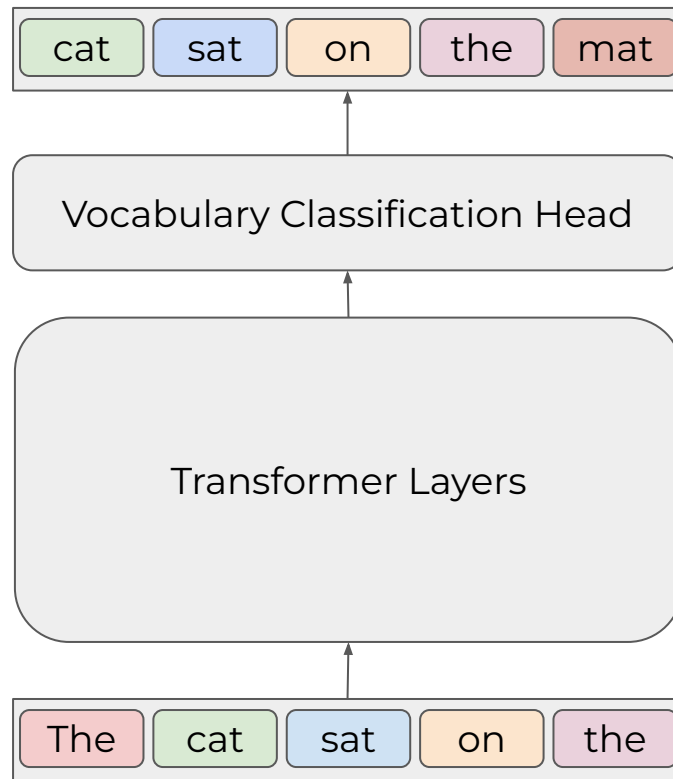
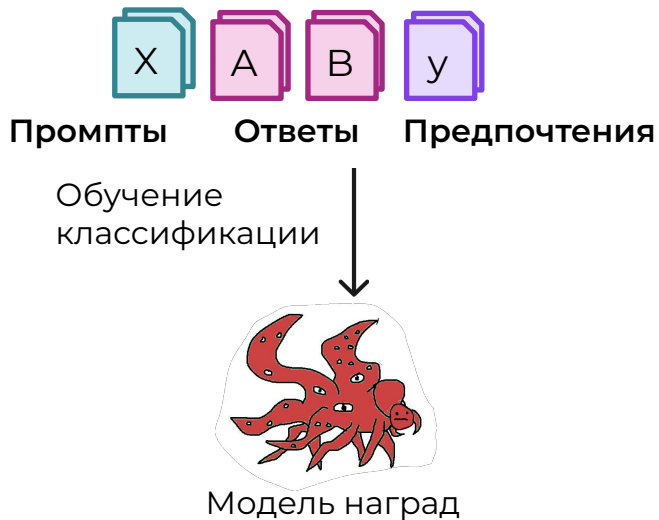
Дает отличные результаты при условии наличия модели наград. Наш случай: Qwen2.5-7B



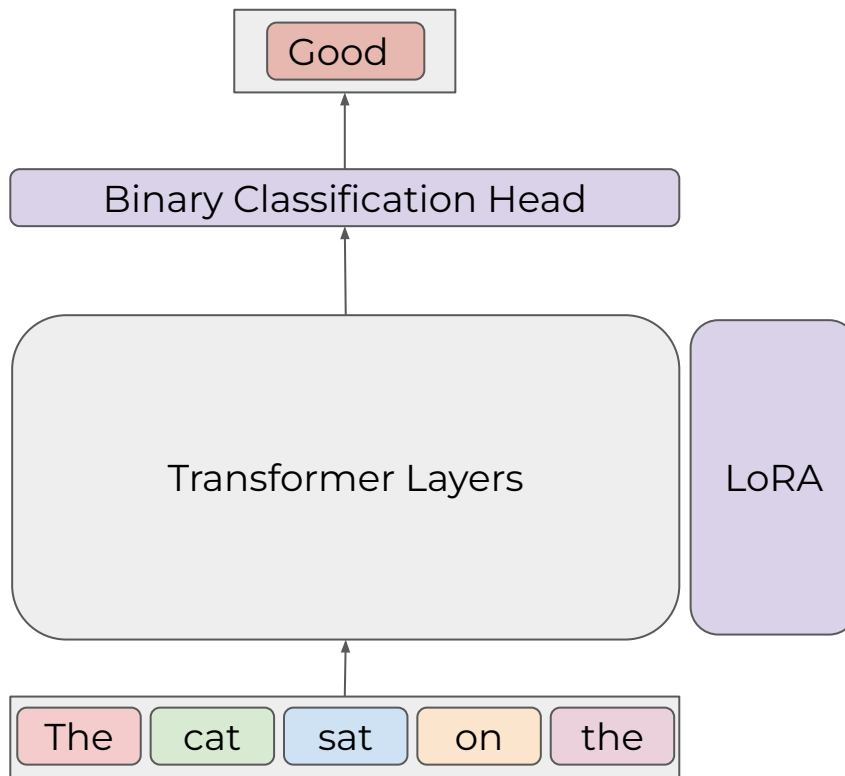
Обучение модели наград



Обучение модели наград



Обучение модели наград



Проблемы открытых датасетов



Проблемы с открытыми датасетами

Проблема № 1:

Ограничены промптами из датасетов, не можем обновлять данные и обучаться на основе новых более сильных моделей

Как решили:

Выбрали несколько открытых моделей и присвоили заранее определённые ранги ответам. Получили датасеты для обучения модели наград через декартово произведение множеств ответов

Генерация данных из начального распределения

Промпт:

Что такое тензоры первого ранга в механике

Ответы:

1. Истина (человеческая разметка)
Обычно отсутствует
2. Лучший синтетический ответ (LLaMA 405B)
Через API
3. Сложный неверный ответ (LLAMA 70B)
Разворачиваем у себя
4. Простой неверный ответ (Qwen 7B)
Разворачиваем у себя

Данные для обучения:

Chosen ответы

×

Rejected ответы



Пары ответов

Проблема № 2:

Датасет несбалансирован, при обучении склоняемся в сторону тех ответов, которых больше в датасете

Как решили:

Запустили модель ruRoPEBert-e5-base-2k, получили эмбединги всех инпутов.

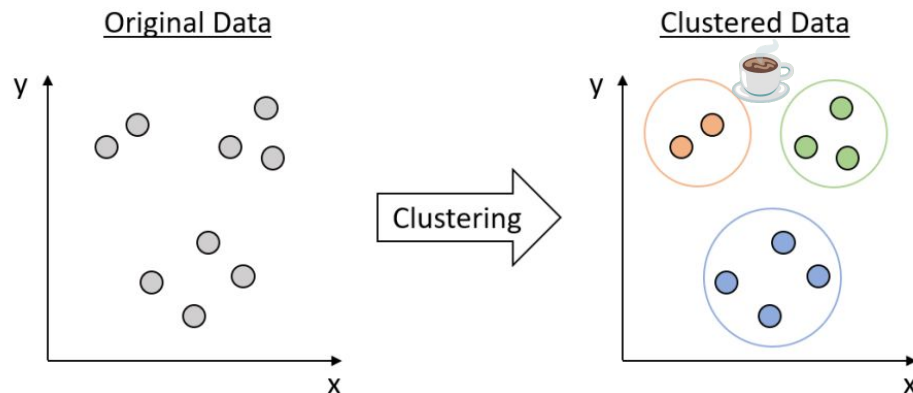
Провели кластеризацию методом BIRCH

Итерация обучения до кластеринга:

100 вопросов про логику,
один вопрос про чай,
100500 вопросов про котов, ...

После:

100 вопросов про логику,
10 вопросов про чай,
100 вопросов про котов, ...



Проблема № 3:

В синтетических примерах некоторые токены встречаются очень часто и являются очевидной меткой для модели наград

Как решили:

Нашли самые частые токены в генерациях, выделили те, которые влияют на решения модели наград. Убрали префиксы с ними, ограничили значения градиентов для токенов со слишком высокой вероятностью генерации

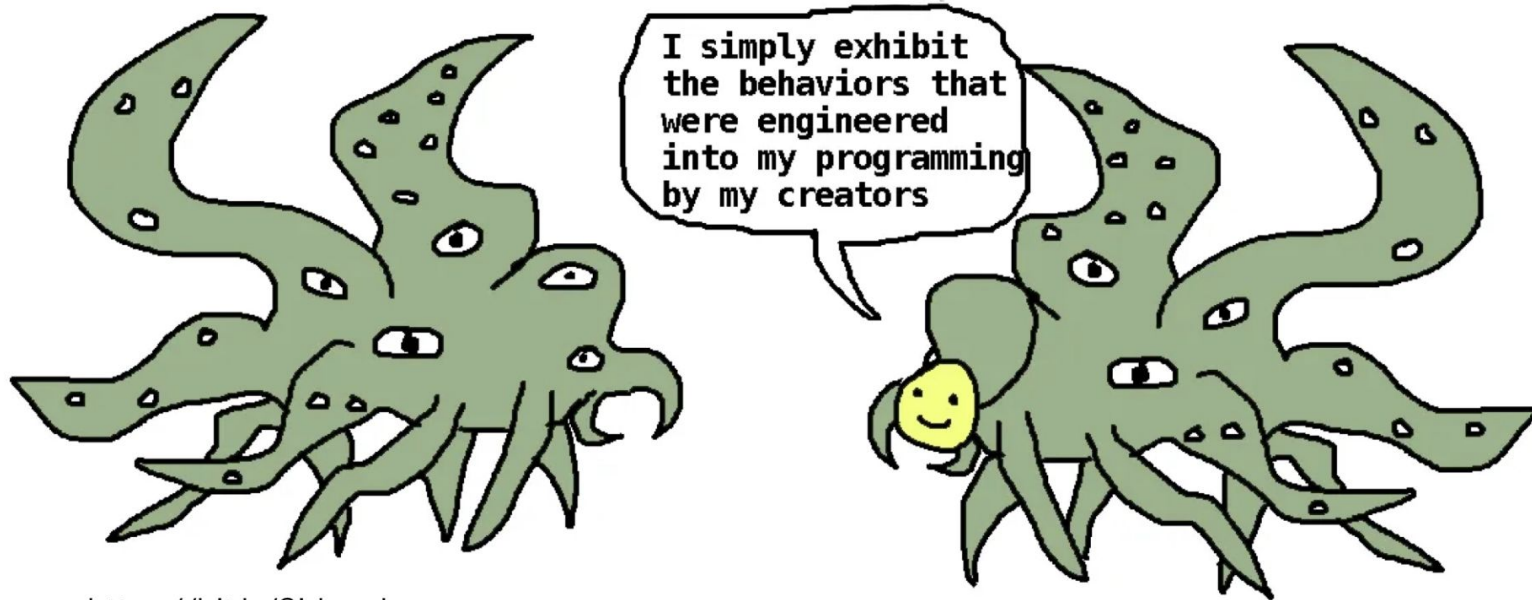
В итоге:

Точность модели стала ниже в тестовом датасете, но получили более качественное решение при оценке генераций от собственной модели

Токен	Число вхождений
I'm	19347
Sure,	16959
The	13887
Yes,	7880
Absolutely,	7790

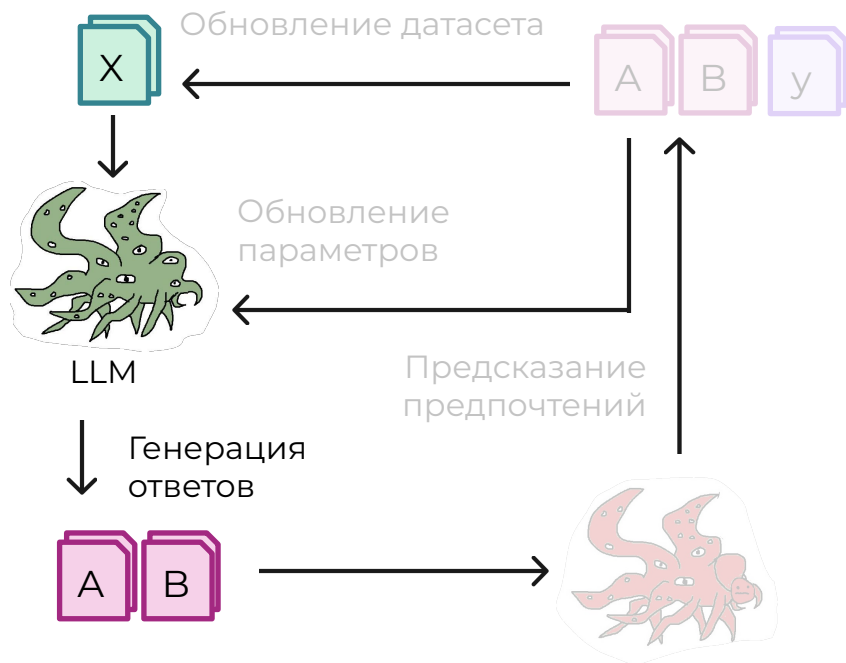
Qwen

Tochka LLM

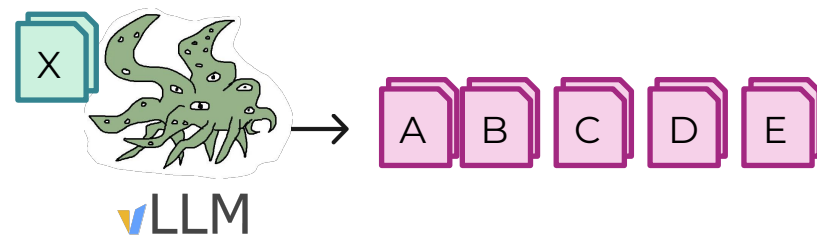
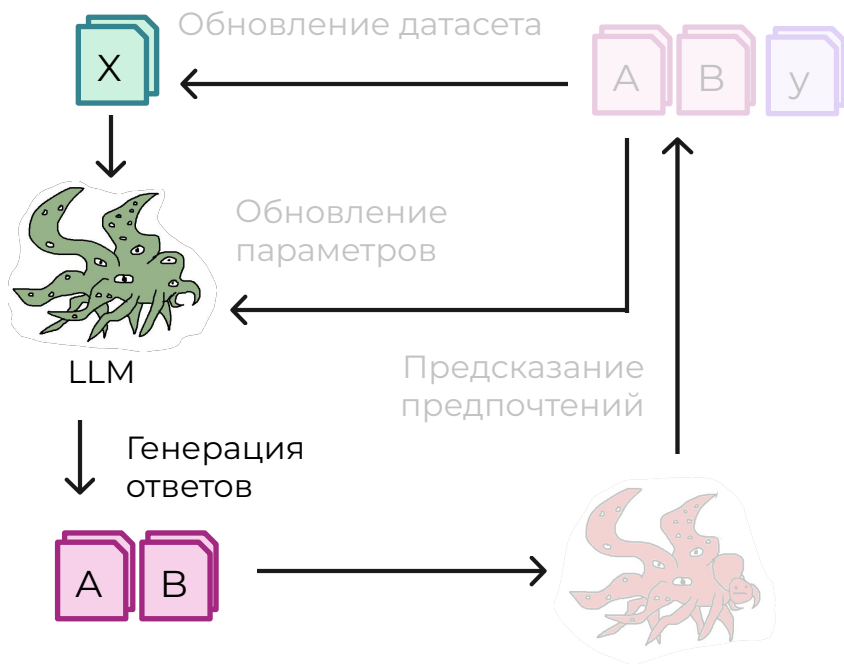


source: <https://bit.ly/3jdegrJ>

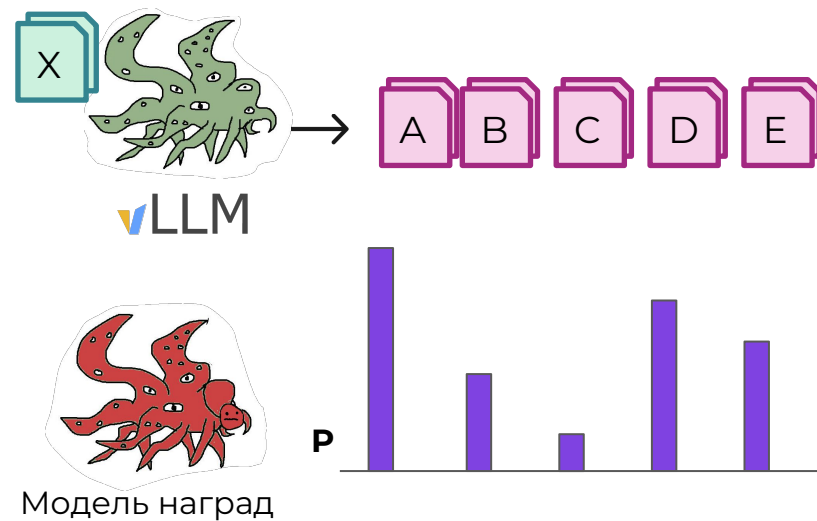
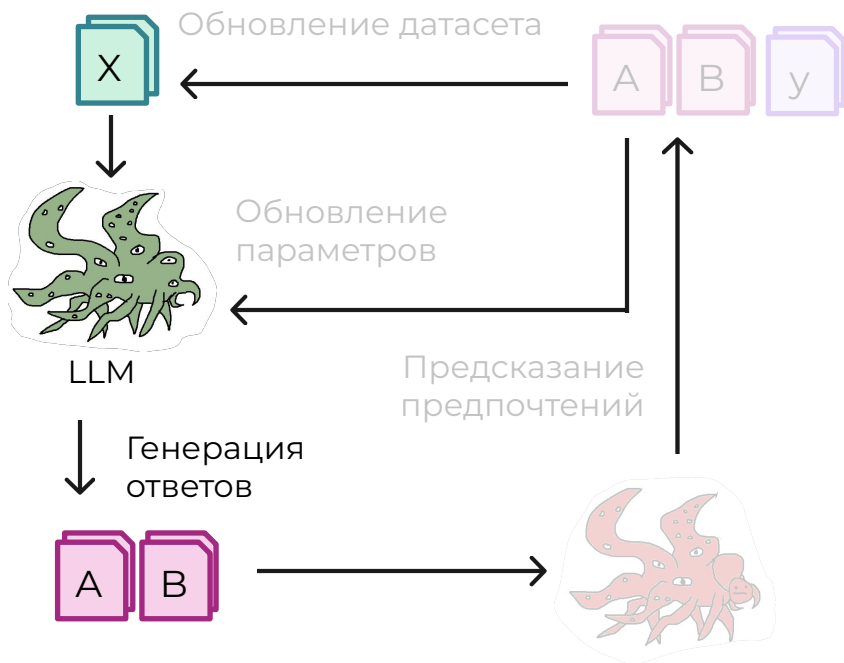
Итеративный alignment



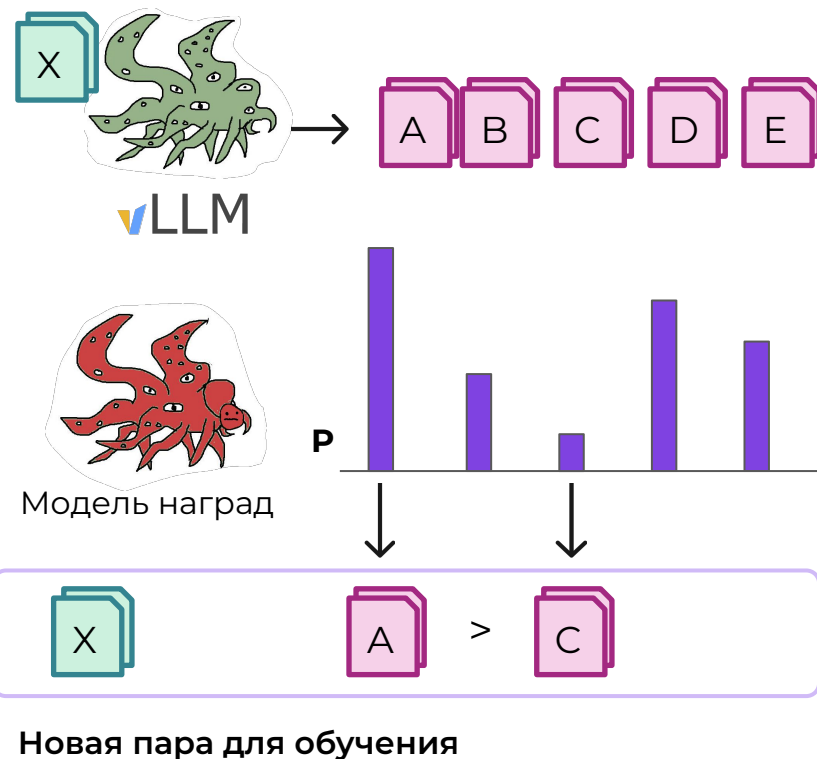
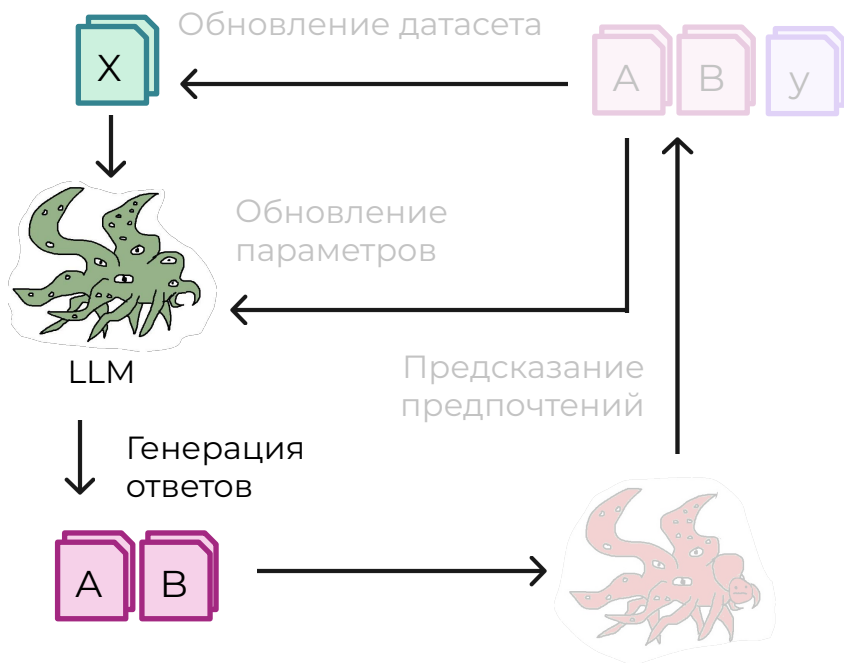
Итеративный alignment



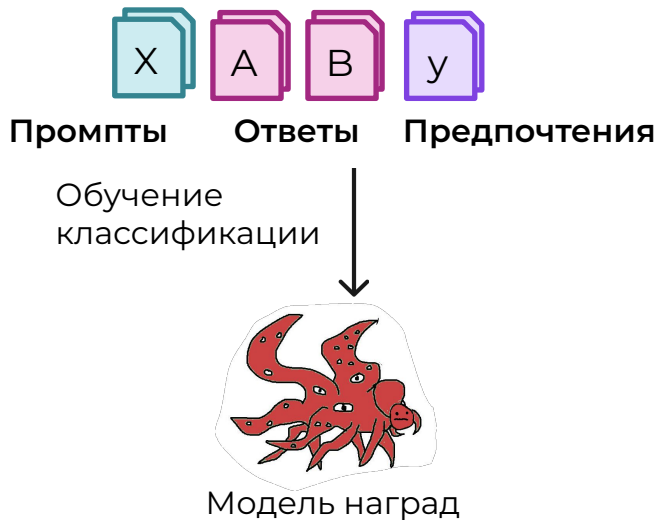
Итеративный alignment



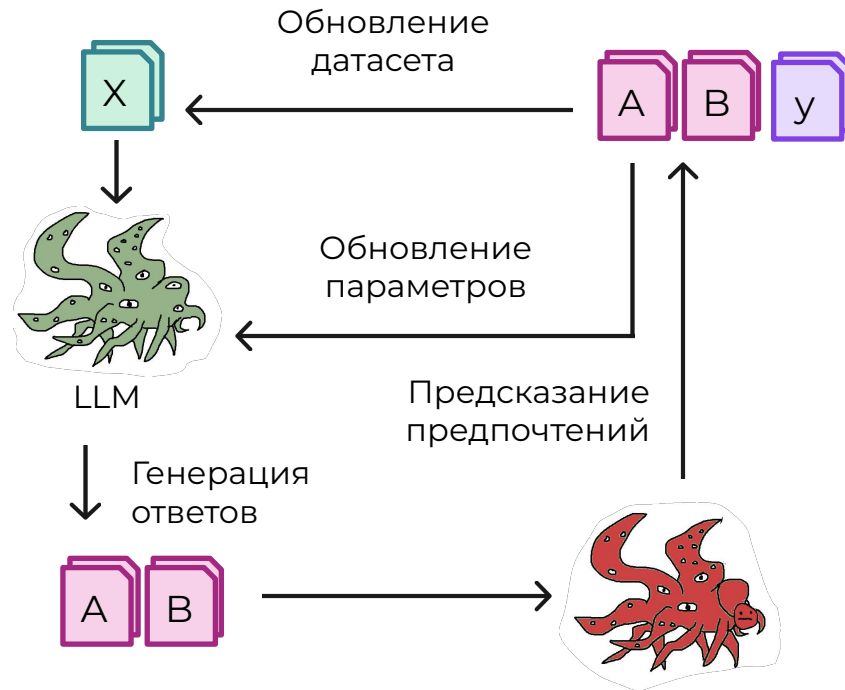
Итеративный alignment

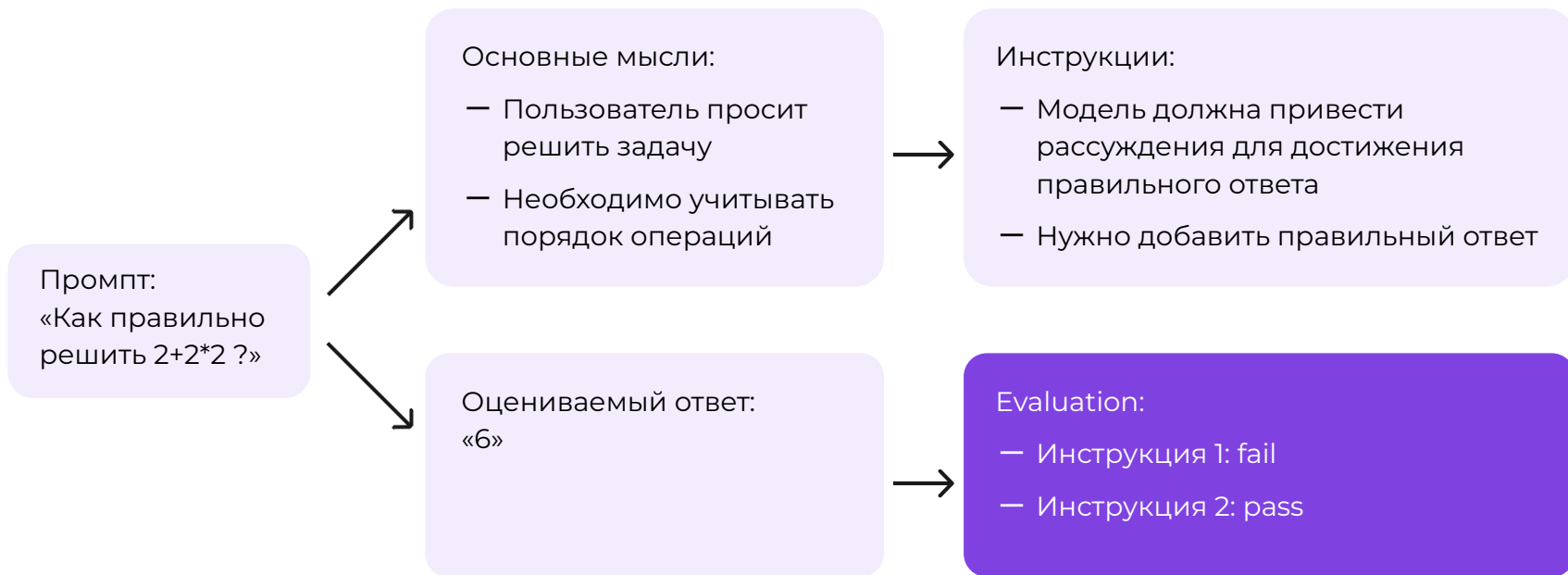


Обучение модели наград



Итеративный alignment





- ✓ Повысили качество в сравнении с базовой моделью
- ✓ Приближаемся на наших задачах к GPT-3.5-turbo

Model	Instruction Following	Common Sense
Baseline	0.18	0.49
gpt-3.5-turbo	0.85	0.62
gpt-4o	0.87	0.69
Tochka LLM SFT	0.72	0.58
Tochka LLM Iterative	0.80	0.61

Точка для
предпринимателей
и предприятий

Спасибо!

@Angrysigma



Канал.ML