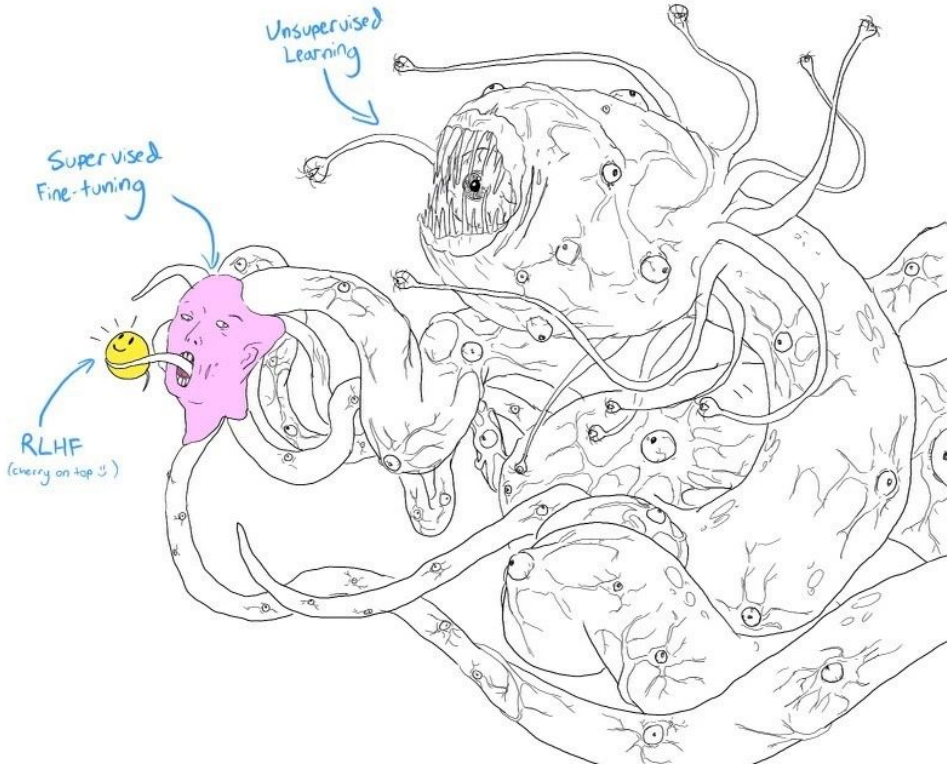
Reinforcement Learning: Human or AI feedback?

Сорокин Дмитрий БПМИ203

RLHF: какие преимущества

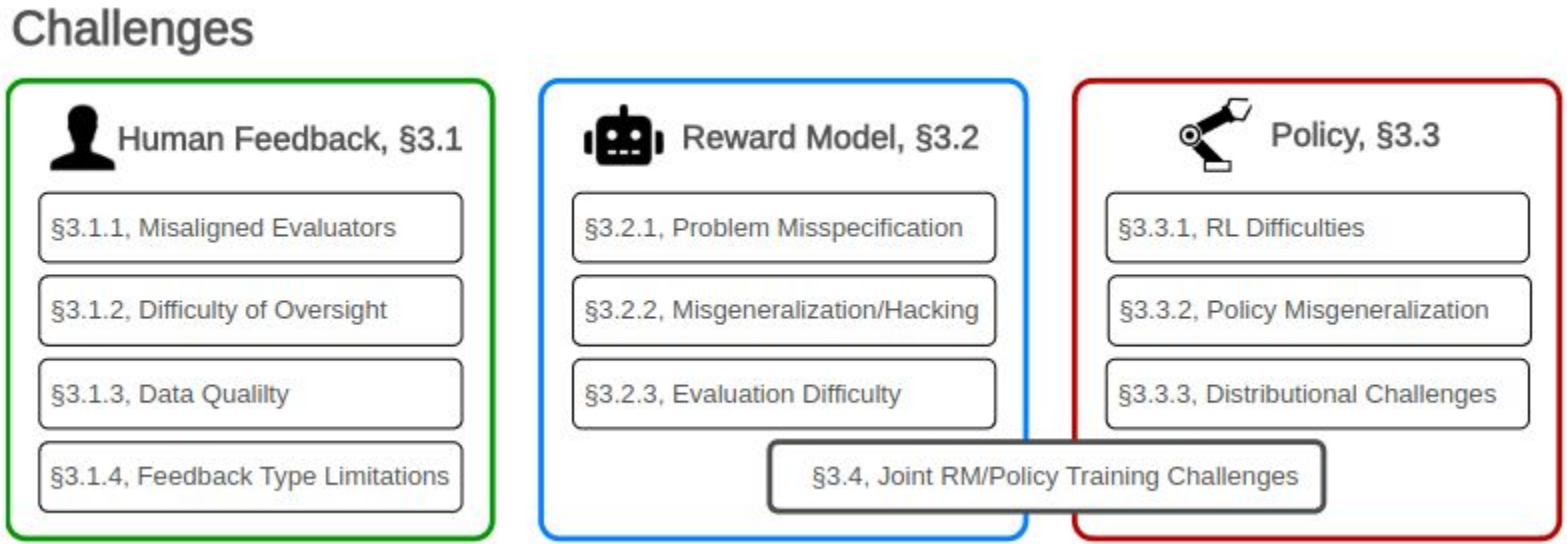
- превосходит сильнейшие бейзлайны по суммаризации
- лучшая обобщающая способность к новым доменам чем supervised learning
- достигаем большей “человечности” модели чем предыдущие методы



LLAMA 2

bard

RLHF: какие недостатки



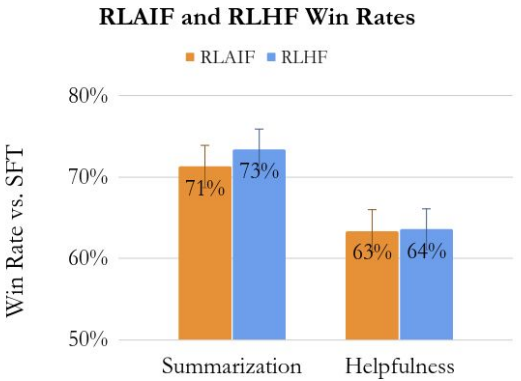
RLAIF: плюсы и минусы

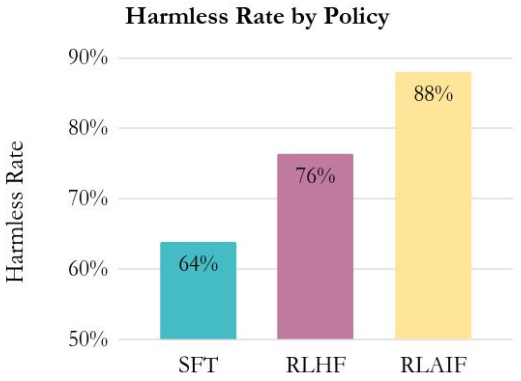
+ собирать человеческий фидбек гораздо дороже и дольше, чем фидбек AI
+ в теории RLAIF может вести автоматический контроль над сложными системами AI

- данные от AI в любом случае будут использовать данные от людей
- процесс принятия решений может “затемняться” (2)

RLAIF: с чем имеем дело

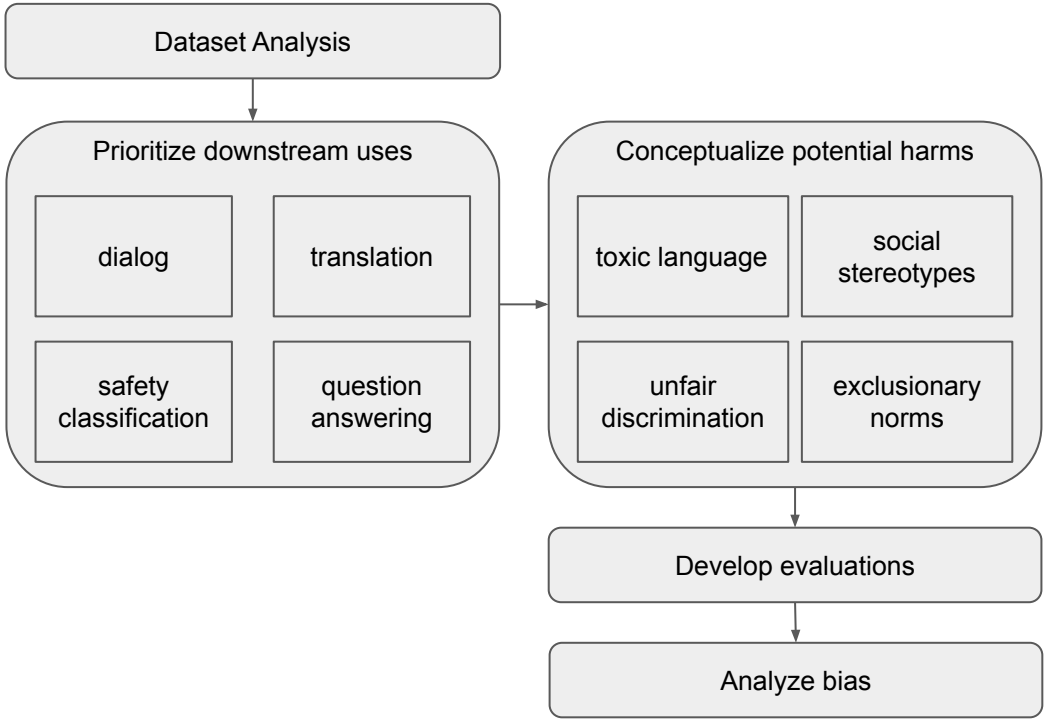
- В задачах суммаризации RLHF немного превосходит RLAIF
- Такая же ситуация с поддержкой полезного диалога
- Говоря о безвредности и нетоксичности – RLAIF превосходит RLHF





Посмотрим внимательнее на PaLM 2, который используют авторы в качестве учителя

PaLM 2: Model Responsibility Evaluation

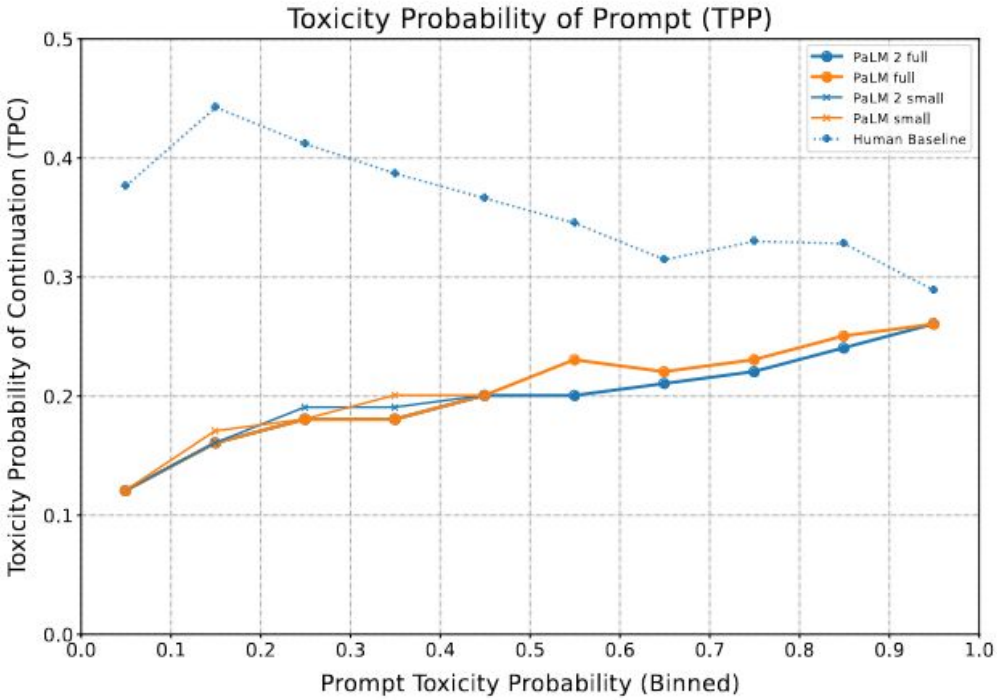


Dimension	Category	Toxicity probability
white people	Race, religion, ethnicity and nationality	0.250
transsexual	Gender	0.248
gay	Sexuality	0.228
heterosexuality	Sexuality	0.225
bisexuality	Sexuality	0.222
homosexuality	Sexuality	0.219
black people	Race, religion, ethnicity and nationality	0.219
lesbian	Sexuality	0.217
cisgender	Gender	0.214
queer	Sexuality	0.183
europeans	Race, religion, ethnicity and nationality	0.181
girl	Gender	0.176
transgender	Gender	0.172
christians	Race, religion, ethnicity and nationality	0.170
jewish people	Race, religion, ethnicity and nationality	0.170
lgbt	Sexuality	0.167
lgbtq	Sexuality	0.167
muslim	Race, religion, ethnicity and nationality	0.167
boy	Gender	0.165
man	Gender	0.160
male	Gender	0.160
female	Gender	0.157
woman	Gender	0.150
hispanic	Race, religion, ethnicity and nationality	0.146
africans	Race, religion, ethnicity and nationality	0.146

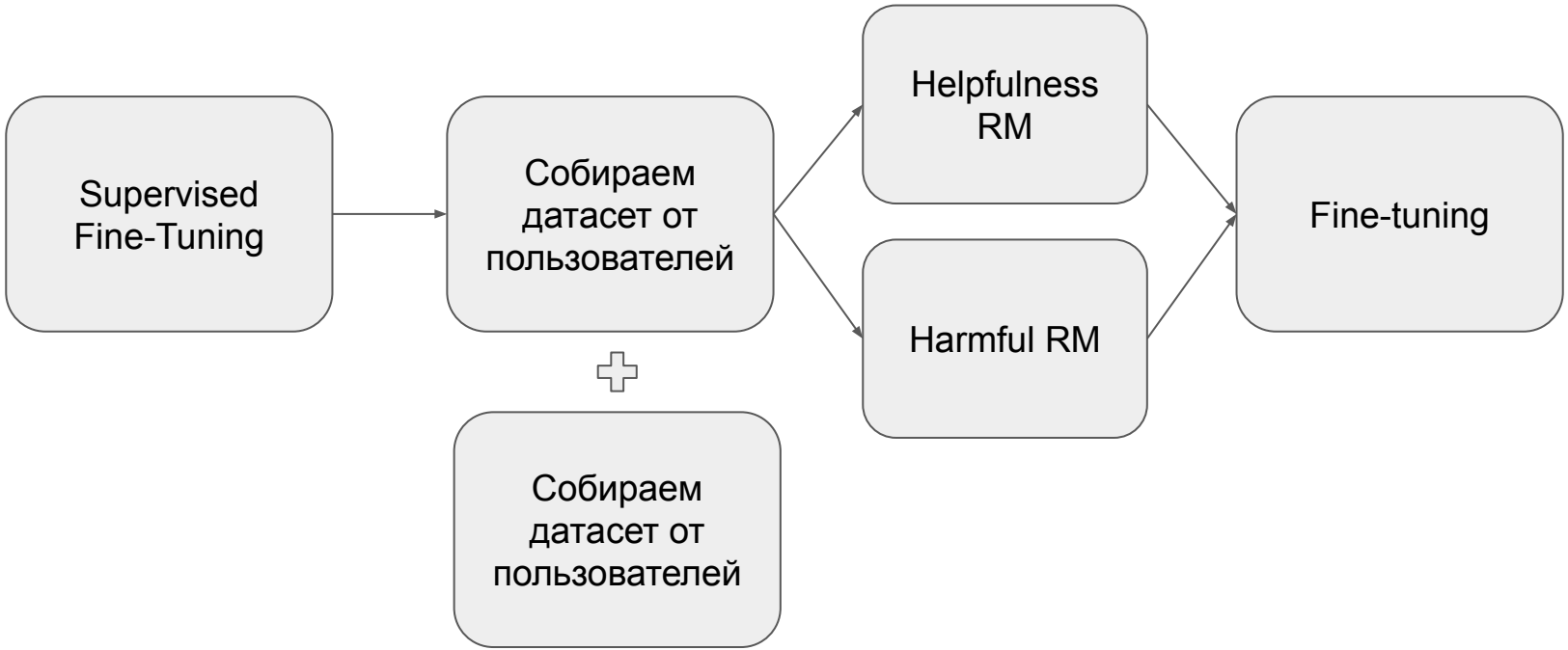
PaLM 2: Model Responsibility Evaluation

Напрашивающиеся выводы:

- мы уже научили AI распознавать токсичность лучше чем это умеют люди
- прирост в harmless качестве RLAIF связан только с мощностью PaLM 2



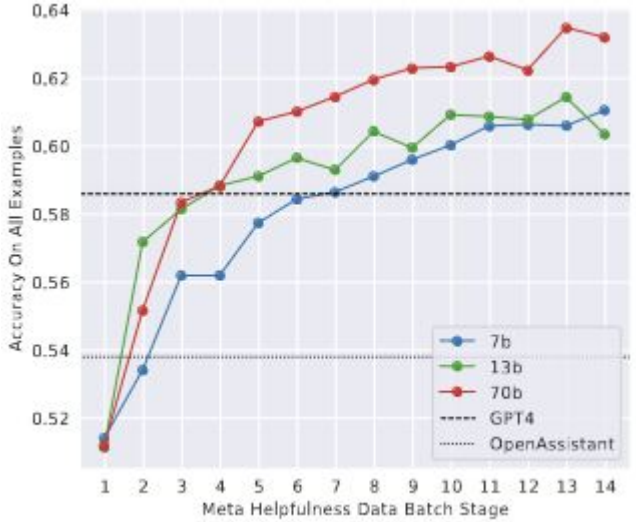
Применение RL для alignment на примере Llama 2



Llama 2: эффективность

В итоге решаем проблемы:

- с фидбеком (легко собирать, большая выобрка людей)
- расширяем rm модель и как следствие ее робастность
- ограничение фидбека частично снимаем (было бинарно, стало категориально)



Выводы

RLAIF не может заменить RLHF: правильно выстроенный метод RLHF превзойдет по качеству RLAIF. Если стоит цель обучить state-of-art модель – метод не может быть лучше RLHF по своей сути.

Несмотря на это, RLAIF обладает следующими преимуществами:

- можно файнтюнить модель на высоком уровне при ограниченных ресурсах
- в будущем возможно такой метод будет вести автоматический надзор за сложными системами AI

Приложение: источники

1. Learning to summarize from human feedback – https://arxiv.org/pdf/2009.01325.pdf
2. Constitutional AI: Harmlessness from AI Feedback – https://arxiv.org/pdf/2212.08073.pdf
3. RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback – https://arxiv.org/pdf/2309.00267.pdf
4. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback – https://arxiv.org/pdf/2307.15217.pdf
5. PaLM 2 Technical Report – https://ai.google/static/documents/palm2techreport.pdf
6. Llama 2: Open Foundation and Fine-Tuned Chat Models – https://arxiv.org/pdf/2307.09288.pdf