1. Вклад статьи
    1) До этого ассистенты могли выдавать либо бесполезную, либо опасную информацию. Это две противоположные между собой характеристики, и задача заключается в том, чтобы найти баланс для них. Данное исследование в свою очередь смогло решить эту проблему
    2) Эта статья довольно сильно опирается на RL из-за этого исследователям пришлось сильно покопаться как это устроено и мы в итоге из статьи видим некоторые новые эффекты, н-р, линейная связь между наградой и корнем КЛ дивергенции
    3) к этой статье прикреплён репозиторий с открытой базой данных взаимодействия модели с человеком. Но всё же воспользоваться ею довольно сложно, о чём я расскажу чуть дальше

2. Авторы

Работа написана в апреле 2022 года командой авторов из лаборатории Anthropic. У статьи порядка 28 авторов, которые разделены на 3 группы по уровню вклада в статью. Три основных автора:

- Yuntao Bai, занимался проектированием экспериментов, а также оценкой качества работы моделей, первостепенная область исследований Yuntao Bai – это физика, и только на втором месте идут NLP исследования в области ассистентов и их безопасности.
- Andy Jones и Kamal Ndoussse, занимались организацией обучения с подкреплением, а также внедрением и отладкой RLHF.
- О вкладе каждого из 28 авторов более подробно можно прочитать в самой статье на стр. 37

3. Лаборатория

Все авторы работают в лаборатории Anthropic, которая специализируется на безопасности, дружелюбности моделей искусственного интеллекта, и у всех авторов общая статья до этой, посвященная языковому ассистенту. Думаю, их следующей задачей было моделирование безопасного ассистента. Специалистов в области RL посетила идея, что в данном случае можно применить обучение на человеческих оценках. Получается, что данная статья - это не случайное открытие, а результат логического развития проектов компании Anthropic

4. Статьи
    1) Ссылаются на:
        - A General Language Assistant as a Laboratory for Alignment – статья, в написании которой участвовали все авторы выше, на неё опирается наша статья
        - Также авторы пользуются наработками в сфере few-shot тестирования,

helpful и harmless формализации и так далее

На основе данной статьи были написаны следующие работы:

- Constitutional AI: Harmlessness from AI Feedback – продолжение статьи, используются AI оценки
- Scaling Laws for Reward Model Overoptimization – более глубокое и детальное изучение эффектов, возникающих во время RL обучения от лаборатории OpenAI
- Inclusive Artificial Intelligence - это такие модели, которые учитывают различные потребности и, следовательно, приносит пользу всему обществу, включая меньшинства, и недостаточно представленные группы. Это достигается за счет уменьшения предвзятости и дискриминации.

2) Конкуренты:

- Teaching language models to support answers with verified quotes (16-21 March 2022) - от DeepMind, в этой статье решается проблема полезности. Здесь также применяется отдельная модель RLHP, но всё же в этой работе модель не всегда безопасна.
- Training language models to follow instructions with human feedback – очень похожая статья от OpenAI. Эта статья станет основой для ChatGPT.
- LaMDA: Language Models for Dialog Applications (10 February 2022)
  Две недавние исследовательские работы LaMDA и InstructGPT были изучены в сравнении с этой работой. LaMDA фокусируется на улучшении диалога на естественном языке за счет внедрения supervised методов обучения, таких как создание и оценка фактов, для создания полезных и безопасных разговоров. InstructGPT фокусируется на использовании обучения с подкреплением для повышения полезности. Наша работа отличается от двух других тем, что в ней основное внимание уделяется «онлайн-обучению», которое включает обучение моделей посредством взаимодействия с человеком для сбора данных более высокого качества. Наконец, в этой работе более подробно рассматриваются различные масштабы и уровни надежности.
- Open-Assistant: некоммерческая организация с членами со всего мира, цель которой — сделать крупномасштабные модели машинного обучения, наборы данных и соответствующий код доступными для широкой публики. Делая модели, наборы данных и код повторно используемыми без необходимости постоянно обучаться с нуля, мы хотим способствовать эффективному использованию энергии и

вычислительных ресурсов для решения проблем, связанных с изменением климата.

5. Сильные и слабые стороны

Ниже приведены сильные стороны данной статьи.

- универсальность и полезность метода
- адаптивность и самосовершенствование(online обучение)
- большое количество иллюстраций в статье

У работы есть и слабые стороны, например:

- слишком большая, новизны знаний нет
- Сложность воспроизведения результатов
- неудобная структуризация текста