TLab.NLP Тестовое 23Q4

Report by Alexey Gorbatovski,

ITMO University, 2023

Introduction

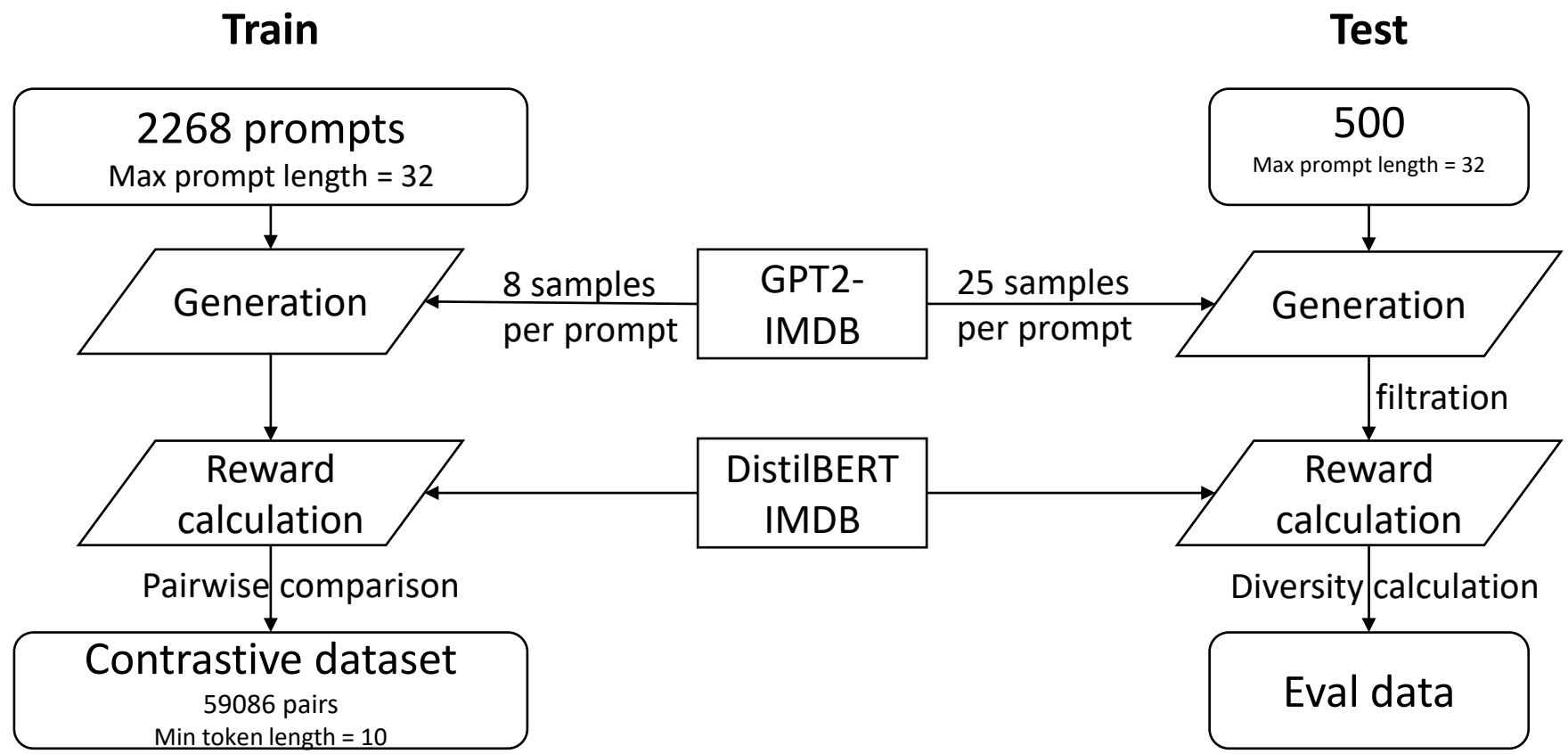
Motivation

- Alignment is crucial technique to adapt LLM for human preferences (HP).
- RLHF is not computationally efficient and stable.
- DPO is the direct way to optimize likelihood of HP.

Methodology and data

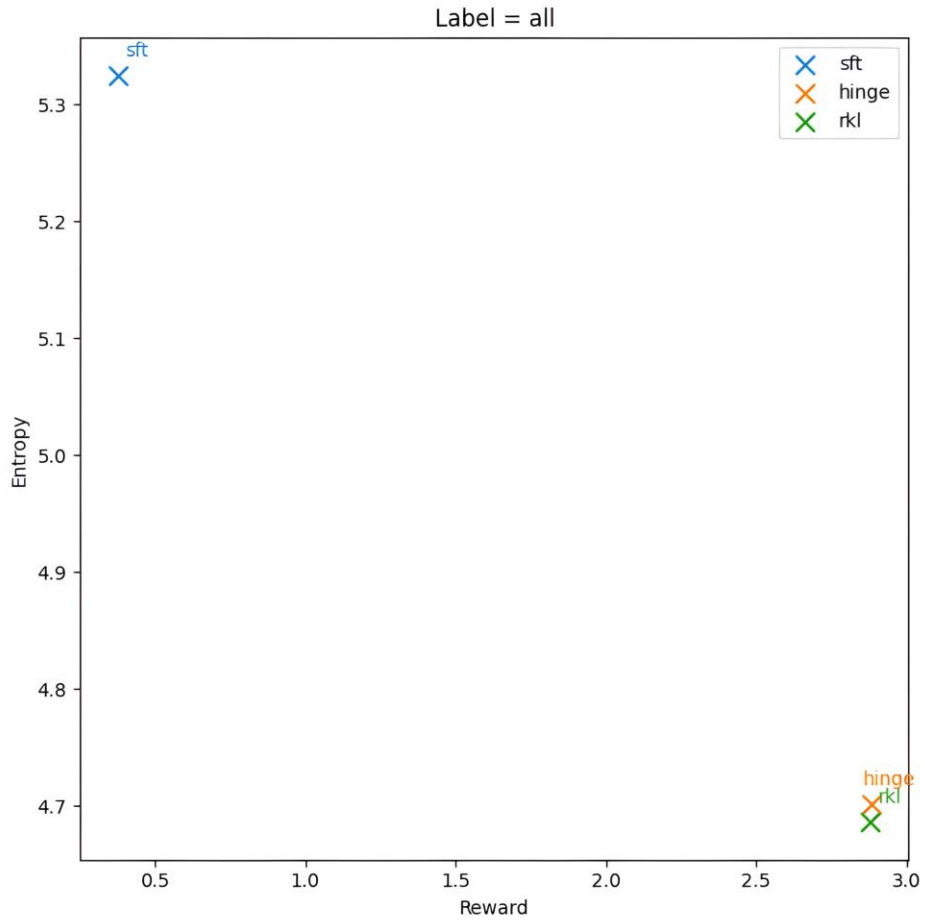
Applied methods:

- DPO (sigmoid/RKL)
- SLiC-HF (hinge)
- F-divergences
- IPO
- cDPO
- Annealing-IPO
- Annealing-DPO

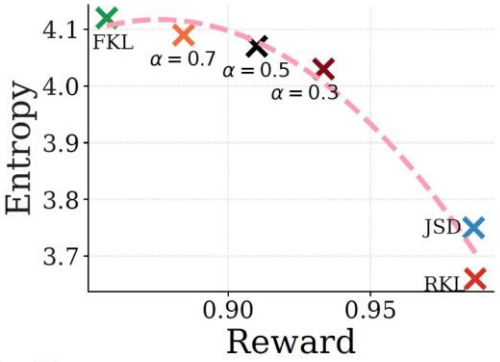


Level 1

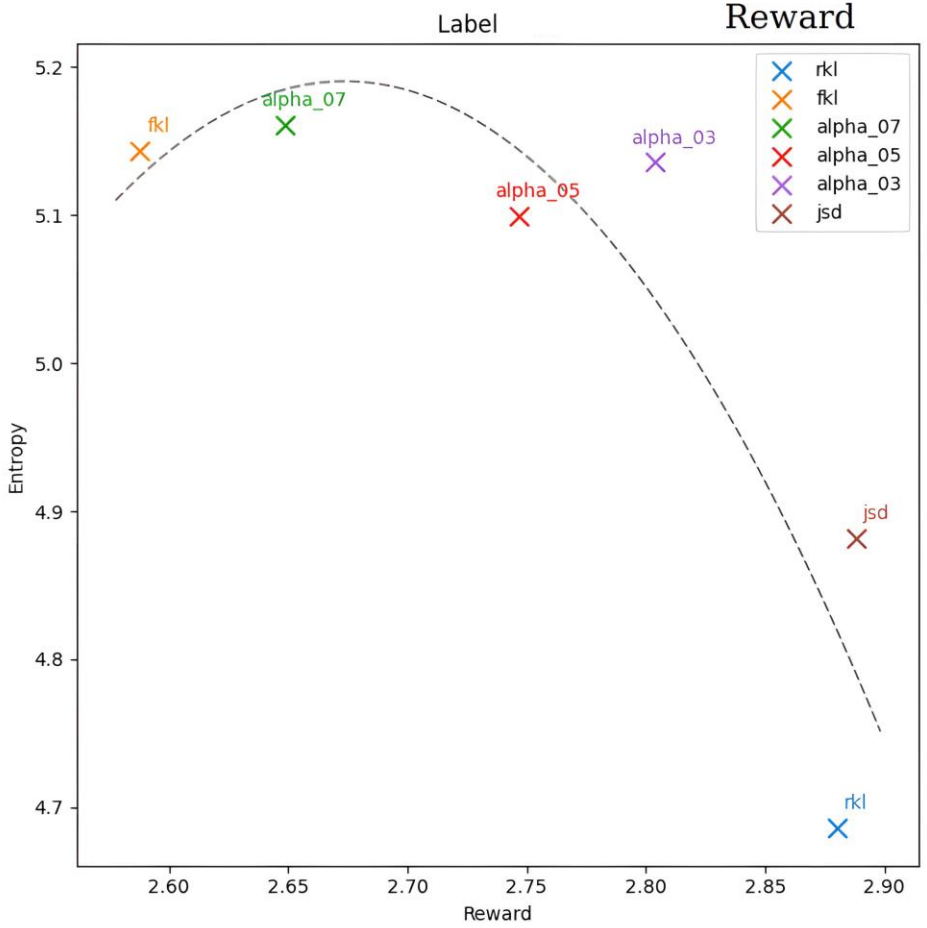
- Базовая модель **SFT** имеет высокий уровень разнообразия при генерации в сравнении с **DPO** и **SLiC-HF**. Так как она не выровнена на определенные предпочтения распределение при генерации более равновероятны.
- **Hinge loss** обучает модель отделять положительные примеры от отрицательных с определенным запасом (margin), в то время как **DPO** (RKL) без регулялизации показывает сильную сходимость к предпочтениям человека.



Level 2 – f-divergences



- **Reverse KL** фокусируется на одной моде распределения, что и обуславливает высокую награду. **JSD** имеет аналогичное поведение, но с легким смещением.
- **Forward KL** дает более равномерное распределение ответов, способствуя сбалансированной политике между разнообразием и наградой.
- **α-дивергенции** оказываются между **FKL** и **RKL**, что предполагает умеренное разнообразие и награду.
- Сравнение с исходной статьей* показывает схожие тенденции, за исключением α=0.5, где разнообразие меньше ожидаемого. Увеличение числа сэмплов с 4 до 25 не привело к изменениям.
- Примечание: расчет энтропии проводился отдельно для каждого промпта, затем агрегировался по группам.



resource* - https://arxiv.org/pdf/2309.16240v1.pdf

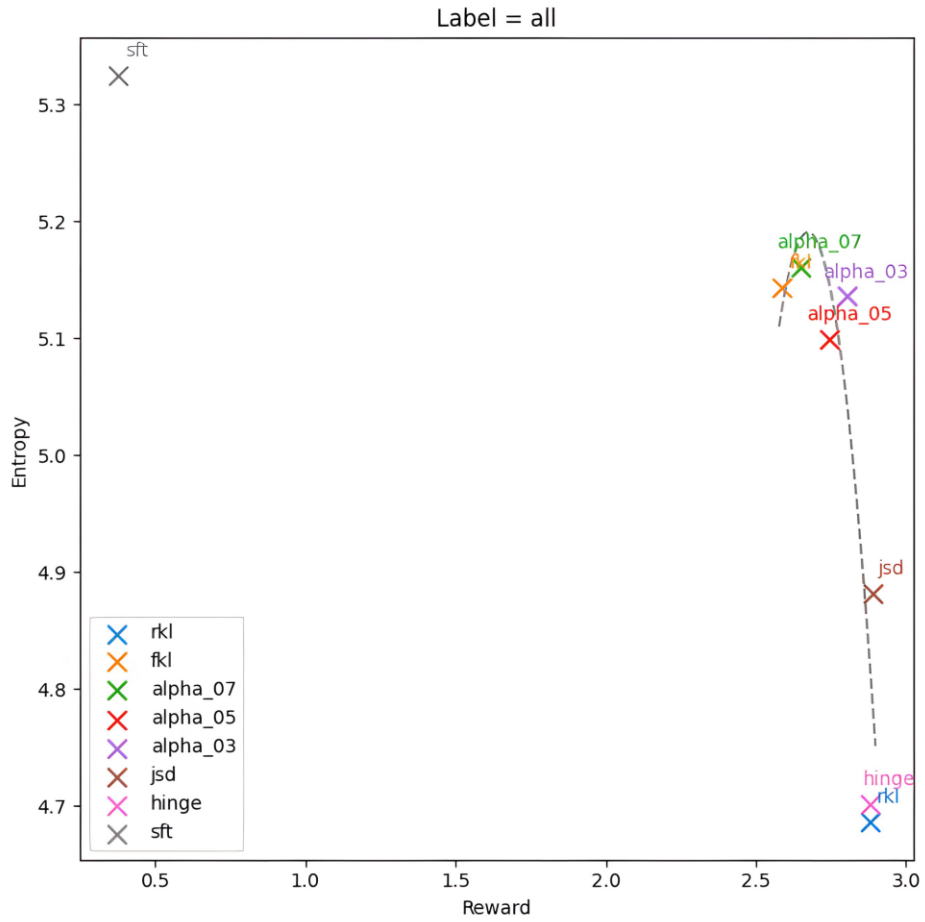
Level 3 - purpose

Цель:

- Хочется чтобы модель имела высокую награду и хорошее разнообразие

Мотивация:

- Intrinsic Dimension у LLM меньше чем у человека*



https://arxiv.org/pdf/2306.04723.pdf *

Level 3 - hypotheses

Использование IPO:

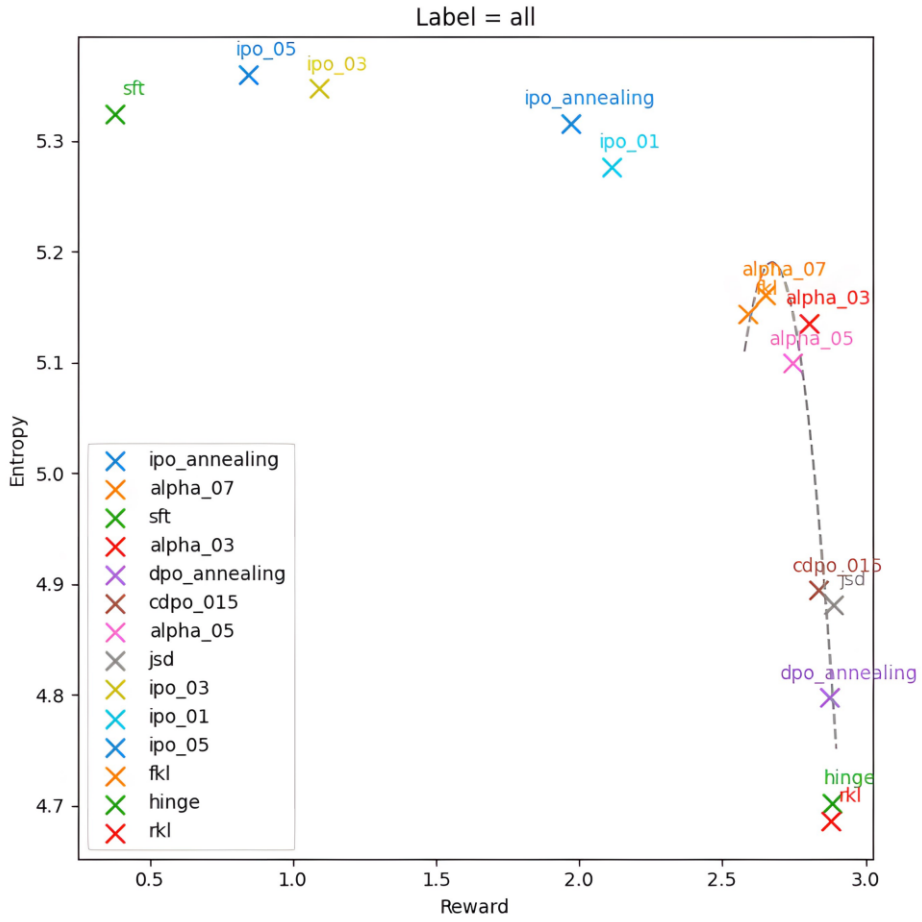
- IPO отклоняется от модели Брэдли-Терри и направлено на решение специфического уравнения, что помогает снизить переобучение на наградах. Для β -> 0 политика будет более вырожденная.
- В ходе экспериментов с различными значениями β была обнаружена прямая зависимость: чем ближе β к единице, тем выше энтропия и ниже награда, что указывает на уменьшение переобучения политики.

Использование cDPO (Conservative DPO):

- cDPO увеличивает разнообразие генераций и уменьшает переобучение на человеческих предпочтениях, заменяя модель Брэдли-Терри на Cross Entropy с небольшим значением ε. Это предполагает вероятность, что вариант с проигрышем может быть лучше.
- Применение cDPO с ε равным 15 привело к созданию более разнообразной модели без значительной потери в наградах.

Обучение с изменением β (annealing) :

- Последовательное изменение β может оптимизировать обучение для генерации как разнообразных, так и соответствующих предпочтениям образцов. Цель состоит в том, чтобы модель изначально вела себя случайно и постепенно концентрировала внимание на эталонной политике. Уместны аналогии с exploration / exploitation. Также, так как первоначальный RL objective схож с ELBO в VAE – возможно интерпретировать это как в работе beta-VAE*. Однако ожидается, что для IPO и DPO будут наблюдаться различные результаты: DPO склонно к сильному сходству к оптимальной награде**, в то время как IPO может сбалансировать энтропию и награду.
- Предположительно гипотеза подтверждается, поскольку **DPO Annealing** показывает меньшее разнообразие по сравнению с **IPO Annealing**. С учётом использования Cosine LR scheduler, среднее значение β в процессе обучения должно было составить около 0.69, что не соответствует наблюдаемой общей тенденции для β в IPO и можно говорить что модель могла выучить разнообразную оптимальную политику.

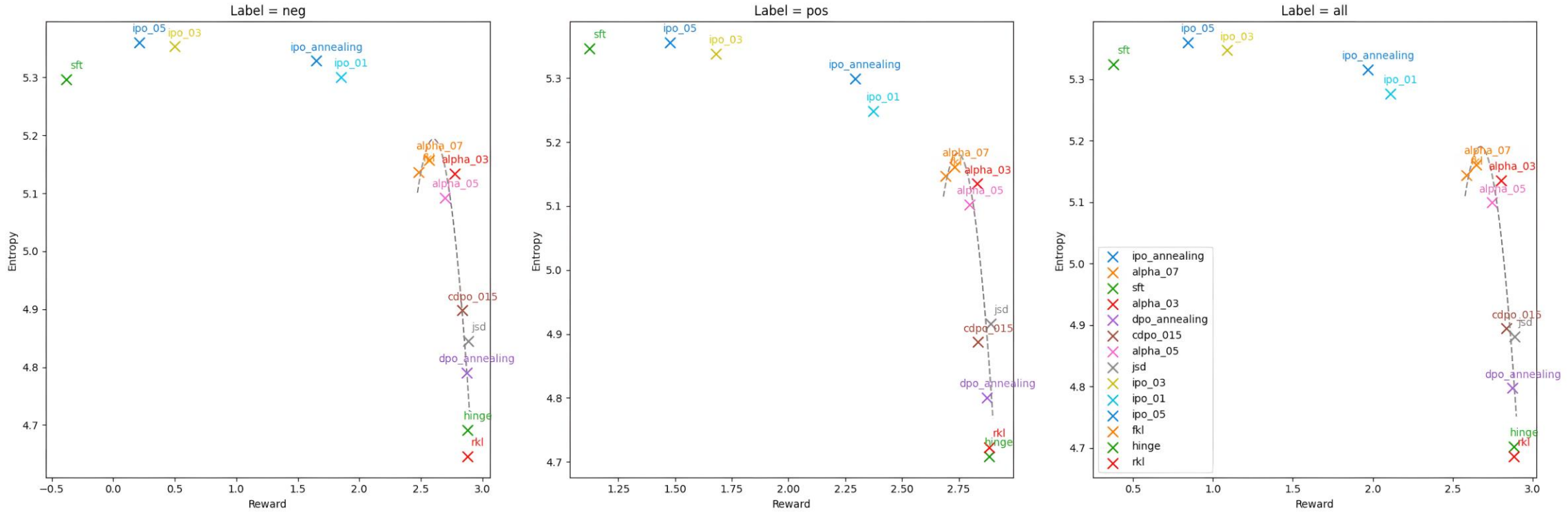


https://openreview.net/pdf?id=Sy2fzU9gl *

https://arxiv.org/abs/2310.12036 **

Original label bias

- Есть зависимость награды от первоначальной метки label в тестовой выборке, возможно, это происходит из-за того, что в 32 токенах prompt протекает информация о sentiment контексте отзыва.



Conclusion

- DPO точно подвержено переобучению, что ведет к меньшему разнообразию в генерации
- Техники IPO и cDPO добавляют регулярицию к изначальному objective DPO, избавляя от переобучения, однако требуется более детальное изучение этого вопроса, как и вариантов выбора beta (прим. annealing).
- Возможно добавление Adversarial Loss может улучшить разнообразие генерируемых ответов и оставаться оптимизированной под предпочтения человека (APO*).

https://arxiv.org/abs/2311.08045 *