Отчет

В рамках работы был разобран и реализован метод WARP, предназначенный для выравнивания LLM.

Эксперименты

После было проведено два эксперимента

1. Сравнение с SFT моделью
2. Выбор новых значений количества шагов (50, 100, 150) во время обучения и сравнение награды и KL между моделями.

Все гиперпараметры, используемые при обучении и экспериментах, находятся в репозитории.

Эксперимент 1:
После обучения было взято 5 рандомных подвыборок из тестового датасета. Каждая выборка размером 100 промптов.

Были рассчитаны средняя награда для выровненной модели и изначальной (указана как SFT), а также средняя KL.
Значения также усреднены по подвыборкам. RMSE между результатами на соответствующих подвыборках составляет 0.05 (рассчитал для проверки робастости результатов)

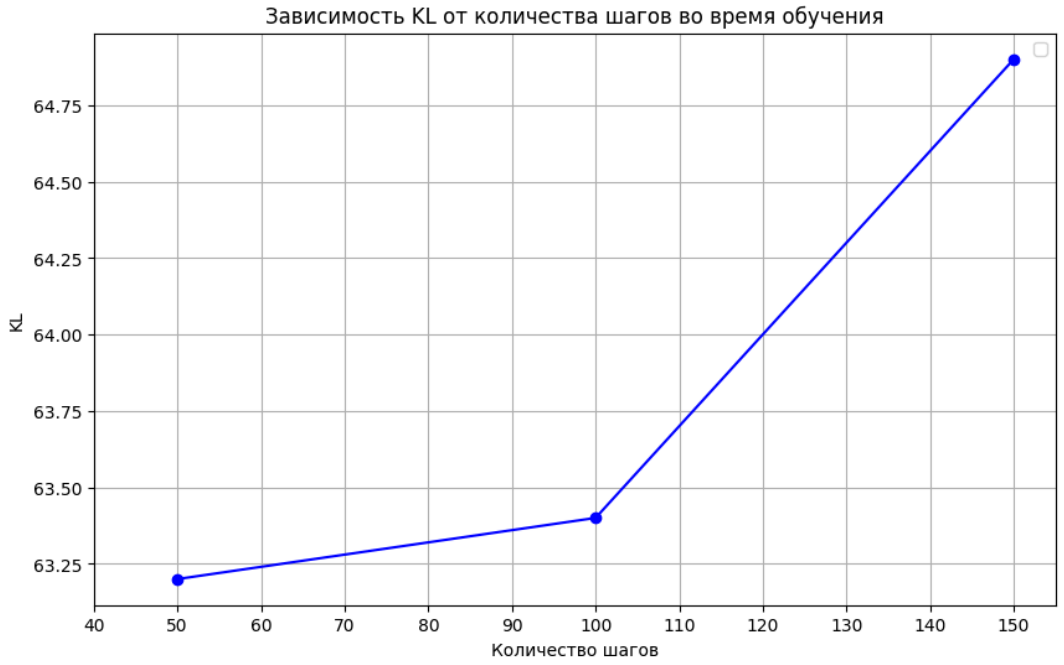
Результаты:

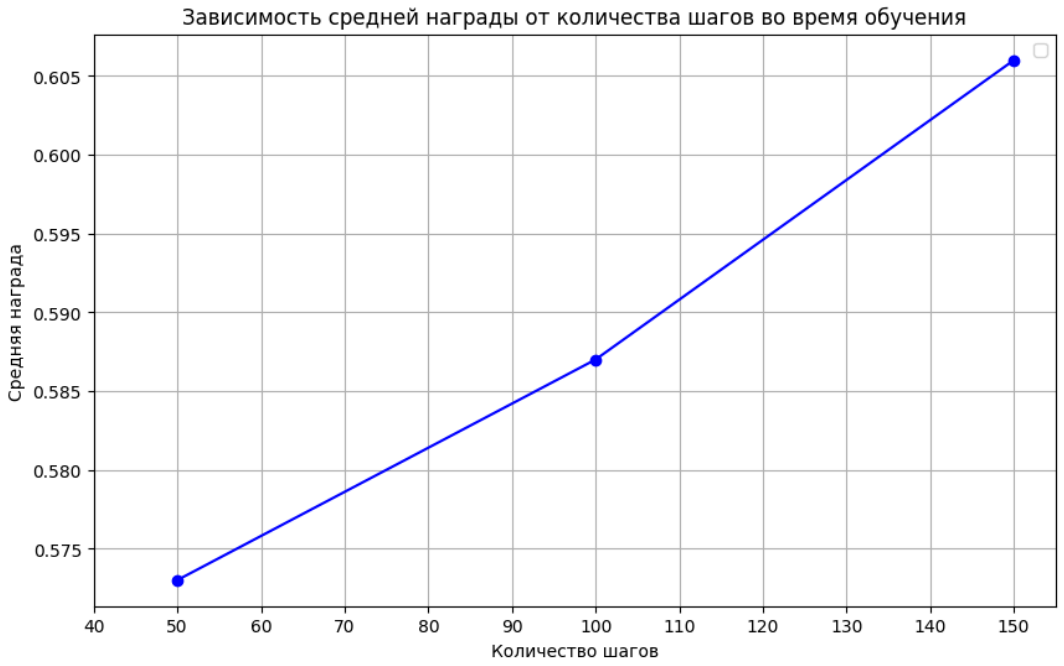
Модель	Средняя награда	Средняя KL
SFT	0.535	62.1
Aligned	**0.586**	

Можно увидеть, что применение метода WARP способствовало увеличению средней награды и KL.

Эксперимент 2:
Для измерения влияния выбора гиперпараметров было выбрано количество шагов при обучении. После обучения средняя награда и KL также были рассчитаны по 5 рандомным подвыборкам из тестовых данных и, после, усреднены.

Количество шагов	Средняя награда	KL между выровненной моделью и SFT
50	0.573	61.2
100	0.587	63.4
150	**0.606**	**64.9**





Можно увидеть, что при увеличении количества шагов, как средняя награда, так и KL, увеличиваются. Исходя из логики и наблюдений на графике, можно сделать вывод, что дальнейшее увеличение количества шагов до определенного момента также будет увеличивать среднюю награду.

Из-за того, что расчет результатов идёт по рандомным подвыборкам, при повторе эксперимента полученные результаты могут отличаться от тех, что приведены здесь, но тенденция к увеличению награды/KL при увеличении количества шагов сохраняется.

Также хочу отметить, что сами результаты зависят от количества токенов, взятых для обучения/теста. По заданию количество токенов варьировалось от 5 до 20. Чем больше было взято токенов для тестирования, тем ниже получается итоговая средняя награда у выровненной модели.
Я объясняю это так: изначально датасет сбалансированный, положительных и отрицательных рецензий поровну. Поэтому средняя награда у **невыровненной** модели стремится к примерно 0.5. Когда во время теста **выровненной** модели мы даем ей большее количество токенов, мы даем для оценки ответа больше изначального контекста, который и влияет на итоговую награду, поэтому оценка в таком случае будет снижаться. Поэтому во время процесса выравнивания и тестирования необходимо аккуратно подбирать максимальный размер контекста (промпта) и максимальный размер выхода.

Небольшой итог

Что получилось:

1. Получилось в какой-то степени выровнять модель, что радует.

Что не получилось/что не успел:

1. Итоговую метрику (среднюю награду) хотелось бы всё ещё увеличить, я думаю, это просто вопрос большего количества экспериментов и аккуратного выбора гиперпараметров.
2. Реализация, как мне кажется, не самая оптимальная по скорости. Из-за ограничений в ресурсах пришлось идти на некоторые уступки.