Дедлайн: 11 августа, до конца дня.
Отправить результаты: @straid, i.pershin@innopolis.ru.

Тестовое задание основано на изучении статьи “Direct Preference Optimization: Your Language Model is Secretly a Reward Model” и состоит из двух частей.

1. Теоретическая часть.
   Напишите краткое изложение статьи своими словами (1 страница A4), отразив
   - основную проблему, рассматриваемую в работе;
   - ключевые идеи и подход DPO;
   - преимущества и недостатки по сравнению с подходами, основанными на обучении с подкреплением. В обучении с использованием DPO обычно используются пары ответов (chosen/rejected), полученные от другой модели или более ранней версии той же модели. Обсудите преимущества и недостатки такой off-policy постановки задачи: какие проблемы могут возникнуть, когда модель обучается на ответах, которые она сама не генерировала?
2. Практическая часть.
   - Вручную реализуйте DPO, без использования готовых реализаций (например, trl.DPOTrainer). Допустимо использовать HF transformers, datasets, peft, accelerate и другие вспомогательные библиотеки, но обучение и логика DPO должны быть реализованы самостоятельно.
   - Используйте небольшую модель (например, GPT-2) и ограниченный набор данных (например, 3000 примеров из Anthropic HH-RLHF).
   - Проанализируйте изменения поведения модели после обучения.

Оценивается:
- глубина понимания DPO;
- читабельность и структура кода;
- качество анализа результатов;
- качество упаковки решения (инструкция по воспроизведению результатов, Dockerfile, requirements.txt и пр.);
- (опционально) эффективность fine-tuning с помощью PEFT или других техник.

Ожидается:
- 1 страница A4 с решением теоретической части;
- описание практической части (ограничений на размер нет, но приветствуемся емкое описание);
- ссылка на публичный репозиторий с кодом.