**Тестовое задание для стажеров в команду Data Science IDP**

Задание — написать решение по извлечению сущностей из документов (новостных текстов). Выполните задание в Jupyter Notebook. Ожидаемый результат: ipynb-файл с решением и всеми выводами ячеек, csv-файл с предсказаниями модели залитый в репоизторий на github.

Вам будет дано задание обучить VLM отвечать на вопросы по графикам.

Для этого вам потребуется пройти несколько шагов.

1. Выбрать существующую VLM, при выборе ориентируйтесь на:

- кол-во и качество релевантных данных на претрейне

- размер модели

- простоту в дообучении

- дополнительно можете оценить простоту последующей квантизации и инференса модели

Дайте теоретическое введние в то, как устроены vlm,

- что входит в архитектуру

- как происходит обучение подобных моделей

Вам нужно будет объяснить причину выбора модели, чем аргументированней будет ваш выбор - тем лучше, опирайтесь на пункты по которым вы выбирали модель.

2. Выберите датасет из huggingface (или kaggle, нужно чтобы у него было описание, при желании можете собрать самостоятельно) для задачи chart question answering. Вот несколько примеров:

- https://huggingface.co/datasets/ChartMimic/ChartMimic

- https://huggingface.co/datasets/listen2you002/ChartLlama-Dataset

Подробно аргументируйте выбор этого сета. Обратите внимания как этот датасет может быть связан с теми, на которых модель уже обучалась.

3. Вам нужно будет дообучить модель на этих данных. Помните, что есть разные стратегии дообучения. (для обучения можно использовать Kaggle notebook /colab)

4. Выберите метрику для вашей задачи и объясните ее выбор. Измерьте качество модели до и после дообучения. Сделайте выводы.

В результате будет оцениваться качество проведенного исследования, аргументы и итоговые выводы.