Отчет о работе команды «ПомидоРТы»

1. Описание предлагаемого решения.

Нами предложено следующее решение проблемы: запустить предобученную лингвистическую модель, ее дообучить на основе открытого датасета, изучить изменения качества на различных метриках, выбрать лучшую из них и

За первую половину времени для выполнения цели хакатона были поставлены задачи:

1. Найти предобученную лингвистическую модель и запустить ее.
2. Дообучить модель на основе открытого датасета.
3. Изучить изменения качества на различных метриках, и выбрать ту которая лучше всего отражает релевантность ответа.
4. Применить методы борьбы с переобучением и проверить их уместность.
5. С помощью пределанных исследований обучить модель наиболее эффективно.
6. Добавить к лингвистической модели распознавание голоса, тем самым сделав голосового помощника
7. Создать телеграмм бот, в котором пользователь сможет пользоваться этой моделью.

2. Описание промежуточных результатов.

Нами были выбраны модели Gemma 2 (9 млрд. параметров).

Модель была предварительно обучена на всем датасете miracl/miracl-corpus.

Получены графики, которые показали сходимость loss на валидационной выборке и взрыв градиента на обучающей.

3.Вопросы организаторам хакатона.

1) Выбранное нами направление решение правильное, или мы неправильно понимаем условия задачи?

2) Что более важно: провести исследование и изучение литературы для выбора наиболее эффективной метрики и решения для избегания переобучение или создание конечного продукта с готовым интерфейсом для использования?

3) Требуется ли учитывать решения проблем с правильным распознаванием естественного голоса человека для корректного перевода запроса в текст?

Очень просим ответить на вопросы (в особенности на первый, как можно скорее).