**Вопросы**

1. Основной вопрос - NLP проекты? Что делал, зачем делал, какие были данные(объем датасета, откуда брали), как учил, как измеряли результат.

2. Работал в команде или один?

3. Как организовывали работу над ML проектами.

4. Знание математических алгоритмов machine learning, основные виды моделей и понимаете критерии оценки их работы;

5. NLP (что делали, какие модели, какие данные, зачем делали и какие были результаты);

6. Опыт разработки компонентов диалоговых систем;

7. Опыт работы с генеративными трансформерами(pretrain, finetuning, LoRA, Adapters). Это самое важное для нас - ключевые слова: LLM, GPT-j, Llama, Falcon, Vicuna, MPT, Zephyr, Mistral;

8. Уровень знания и свободного использования библиотек для машинного обучения (Scikit-learn, Pandas, Matplotlib, Numpy, Scipy, Sklearn);

9. Знание Deep Learning фреймворка PyTorch, transformers, pytorch-lightning; Умение выстроить pipeline самостоятельно

10. Software Engineering в контексте ML:

1) Писал ли сервисы вокруг моделей, как они работали - offline(пайплайны) или online(API для моделей)

2) Знание Python

3) Уверенное пользование OS Linux/Unix и системы управления версиями git;

4) Опыт работы с технологиями виртуализации (Docker и пр.);

5) Опыт работы с SQL/NoSQL базами данных;

Из того, что будет + (прокомментируй, пожалуйста, с чем сталкивался и насколько плотно работал)

ONNX, TensorRT, OpenVINO

vLLM, Langchain, TGI, TensorRT LLM

Оптимизация моделей - квантизация, дистиляция, LoRA для LLM Distributed compute: Deepspeed, accelerate, fabric, jax Weight&Biases, DVC, Gitlab CI

Как изучаешь актуальные исследования в области? Что читаешь, статьи и исследования, может быть сам что-то писал(Это большой плюс)

open-source проекты(если куда-то контрибьютишь - тоже большой плюс и лайк)

1.

- Ivr бот для ответа на клиентские запросы. ( аналитика, подготовка датасетов, моделирование (обучение на гпу, инференс на цпу), деплой в продакшн) логируем пользовательские запросы и операторские диалоги. По офлайн метрикам точность, охват, гармоническое среднее между ними. Бизнес метрики: автоматизация, csi, перезвоны.

- ЧатБот в том же ключе, только другой домен данных.

- Суфлер для оператора (ищет документы релевантные запросам клиента во время диалога с оператором (датасет из диалогов) оценка sbs по разметке (обучение на гпу)

- виртуальный помощник, который ищет данные по документам. Помощь работникам искать информацию на предприятии в области закупок. (Нормативные акты, внутренние документы, законы) сравнительная вычитка на точность ответов, юзер фредли и понятность оценивалась со стороны жюри во время финала хакатона (обучение на гпу и инференс на гпу)

- сервис по формированию персонализированных объявлений. Генерация объявлений на основе картинок и кратких заметок в помощь продавцу под профиль клиента. Решали задачу матчинга, ретеншена в сервисе и увеличения скорости конверсии в сделки. Использовали sbs подход по оценке качества генерируемых объявлений. Оценивали жюри в финале хакатона (длительность 6 мес) (обучение на гпу и инференс на цпу)

- соревнование Kaggle по распознование номеров автомобилей. Офлайн метрика CER, лидерборд

- новостная группа в социальной сети (рекомендательная модель, модель классификации текстов, картинок, текстовый трансформер) оценка по метрикам точности, охвата и гармонической меры во время офлайн измерений и онлайн с помощью лайков, количества просмотров и роста аудитории.

2.

Лето 2022г работал один, до этого втроем, а после вчетвером.

3.

Это основная задача с 2022 года

4.

Классические модели до перцептрона в учебных проектах разбирал до матриц. НЛП задачи решаю с помощью готовых библиотек. С теорией механики работы знаком.

5.

- Задача классификации:

Клиентские запросы, чтобы помогать людям, решая проблемы с продуктами компании

- Задача NER

Поиск препаратов в клиентских запросах чтобы помогать людям, точнее определять потребность клиентов

- Задача генерации контента

Генерация объявлений для сервиса поиска недвижимости. Данные с авито и циан, помноженные чатгпт. Результат готовый сервис, который оценивали жюри по перфомансу

- Задача поиска информации

RAG поиск по базе знаний для помощи оператору

- Дедубликация данных

Удаление похожих постов в группе вк

- Суммаризация данных

Суммаризация дилогов для помощи вычитки инсайтов сценаристам

- Fact-extraction

Выделение важной информации из новостей для прогнозирования направления рынка. Опенсорсные датасеты bbc, reuters

- SFT LLM

Под задачи генерации и/или ответов на вопросы

- Обучение pretrain весов для собственного BERT:

под домен компании разработать веса, которые можно finetune уже под задачу

- Обучение pretrain весов для собственного GPT2:

генерация ответов на основе операторских диалогах на вопросы клиентов

6.

4 года разработки (чатботы, виртуальные помощники, ivr боты, социальные боты)

7. Работал с pretrain, finetuning, LoRA. LoRA и есть адаптер. Еще работал с Saiga. Имею опыт работы с LLM, Llama, Mistral. И еще с GPT3, gigachat (like llama2) и gemini (google)

8.

Постоянно использую в работе: Scikit-learn, Pandas, Matplotlib. Во время учебных проектов Numpy, scipy или каких-то вспомогательных проектах чаще аналитических.

9.

Знаю и работаю над построением пайплайнов PyTorch, transformers в НЛП. pytorch-lightning использовал при построении моделей в deep learning.

10.

Botify - приложение по рекомендациям музыки

Ассистенты в тг, боты в вк, веб приложение для поиска недвижимости, приложение для поиска книг, приложение для очистки данных

11. писал пайпланы и на работе и для своих проектов, как онлайн так и офлайн. Подвоз данных офлайн, онлайн взаимодействие с пользователями с полноценным циклом обновления

AirFlow, FastAPI, Streemlit, Docker

12. уверенное

13. уверенное

14. пользовательское, с документацией

15. работаю с postgress, Hadoop. Немного знаком с oracle, Cassandra, redis

16.

ONNX, TensorRT, OpenVINO – слышал, но не использовал

Langchain использовал для реализации RAG подхода

квантизация, дистиляция, LoRA использую часто для тюнинга моделей и подстраивания под условия быстрого инференса

Deepspeed, accelerate, - использовал другие подходы к ускорению обучения

fabric, Weight&Biases - не знаком

jax – использовал для ускорения инференса ллм на цпу

DVC – работаю с этим

Gitlab CI – в учебных проектах тренировался

Контрибьютил в проект

https://github.com/made-mlops-2022