Программа курса
«ARML: Архитектура ML-систем»

О курсе: 6-дневный курс о том, как организовать полный цикл разработки и внедрения систем машинного обучения и нейронных сетей, включая LLM, и эффективно сопровождать их в промышленных решениях с использованием современных подходов и технологий.

Аудитория:

- ML-инженер или Data Scientist (Middle/Senior)
- Архитектор ПО (Software/Solution Architect)
- Team Lead / Tech Lead в Data Science
- DevOps/MLOps-инженер

Уровень подготовки:

- Опыт программирования на любом языке
- Базовые знания System Design (желательно)

Продолжительность курса: 24 академических часа, 6 дней по 4 ак. часа дистанционно

Содержание программы

1. Этапы ML-проекта и инструменты

- Углубленное изучение жизненного цикла ML-решения: от постановки бизнес-задачи до вывода из эксплуатации
- Роли в команде (DS, DE, DevOps, ML инженер) и их взаимодействие
- Обзор и сравнительный анализ современных инструментов и платформ для каждого этапа
    - **Практическое задание** — декомпозиция реального бизнес-кейса на этапы ML-проекта. Выбор стека технологий под конкретную задачу и обоснование принятых решений.

2. Трекинг экспериментов, хранение артефактов и ML-моделей

- Важность воспроизводимости экспериментов
- Архитектура систем трекинга (MLflow, ClearML)
- Принципы организации хранилищ артефактов и моделей
    - **Практическое задание —** проектирование архитектуры для централизованного хранения и версионирования моделей и артефактов в рамках компании.

3. Сценарии использования ML-моделей и Feature engineering

- Паттерны эксплуатации ML-моделей: онлайн (real-time), пакетный (batch) и потоковый (streaming) инференс
- Проектирование Feature Store: архитектура, компоненты, онлайн и офлайн слои
- Паттерны доставки признаков в production
    - **Практическое задание —** разработка архитектурной схемы для системы, обслуживающей модели в режиме реального времени с использованием Feature Store.

4. Инструменты оркестрации и их использование в ML-проекте

- Глубокий разбор и сравнение оркестраторов: Apache Airflow, Argo Workflows, Mage.AI. Их место в MLOps-пайплайне
- Проектирование отказоустойчивых и масштабируемых ETL/ML-пайплайнов
    - **Практическое задание —** проектирование DAG-а для сложного пайплайна, включающего сбор данных, предобработку, обучение модели и валидацию.

5. Инференс ML-моделей

- Архитектура инференс-сервисов
- Обзор и сравнение инструментов: MLServer, Triton Inference Server, TensorFlow Serving, TorchServe
- Оптимизация моделей для инференса (квантизация, дистилляция)

- **Практическое задание** — проектирование архитектуры высоконагруженного инференс-сервиса с учетом требований к задержке (latency) и пропускной способности (throughput).

6. Сопровождение ML-моделей в production

- Концепция наблюдаемости (observability) для ML-систем
- Мониторинг производительности моделей, дрейфа данных (data drift) и концепта (concept drift)
- Подходы к A/B-тестированию моделей
- ML Security и Data Privacy
    - **Практическое задание** — разработка стратегии мониторинга для развернутой ML-модели, включая определение ключевых метрик и настройку алертов.

7. AutoML в MLOps-конвейере (опционально*)

- Как AutoML-инструменты встраиваются в MLOps-цикл
- Обзор state-of-the-art решений
- Экономический эффект от внедрения AutoML и снижение Time-to-Market
    - **Практическое задание** — проектирование пайплайна с использованием AutoML для автоматического подбора моделей и гиперпараметров.

*** Расширенная версия курса 32 часа**

8. Эксплуатация больших нейронные сетей в продуктивном контуре (опционально*)

- Архитектура систем для обучения и инференса больших моделей (LLM, RAG)
- Методы оптимизации: дистилляция, квантизация, прунинг
- Инфраструктура для работы с GPU
- Особенности эксплуатации нейросетей на edge-устройствах
    - **Практическое задание** — проектирование архитектуры RAG-системы для корпоративной базы знаний.

*** Расширенная версия курса 32 часа**