

## Программа курса

### «AGENT: AI-агенты для оптимизации бизнес-процессов»

**О курсе:** 6-дневный интенсивный курс AI агенты для компаний по проектированию, разработке и внедрению AI-агентов на основе LLM (Llama), многоагентных фреймворков (LangGraph, CrewAI, AutoGen) и интеграции с корпоративными системами. Живые занятия с практикой, сопровождение домашних заданий, итоговый проект под ваш кейс индивидуально или в составе группы + консультации по проекту.

#### Аудитория:

- Разработчики и дата инженеры
- ML/AI-специалистам
- Архитекторам/тимлидам
- ML инженерам

#### Уровень подготовки:

- Опыт программирования на любом языке
- Базовые знания System Design (не обязательный навык)

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

## Содержание программы

### 1. Использование LLM

- Архитектура трансформеров. от BERT до Llama 3
- Autoencoders и принципы работы LLM
- Открытый vs проприетарный стек моделей
- Критерии выбора. размер модели, аппаратные требования, лицензирование
- Бенчмаркинг на целевых задачах
- Обзор инструментов Ollama, LM Studio, OpenWebUI

#### Практическая часть:

- Развертывание моделей через Ollama на облачной инфраструктуре
- Сравнение производительности Llama 3 vs Mistral 7B
- Prompt инжиниринг на открытых моделях под задачи суммаризации встреч, анализа кода, генерации скелета презентации

### 2. RAG для обогащения моделей внешними источниками знаний

- Использование RAG (Retrieval Augmented Generation) для обогащения моделей
- Архитектура RAG системы с opensource компонентами
- Выбор моделей эмбедингов под различные задачи и языки
- Векторные базы данных: In-memory решения (FAISS, Annoy)
- Специализированные векторные БД (Chroma, Qdrant, Weaviate, Milvus/Zilliz)
- Расширения векторных БД PgVector для PostgreSQL
- Критерии выбора. производительность, масштабируемость

#### Практическая часть:

- Построение RAG с использованием Sentence Transformers
- Загрузка данных в Qdrant
- Настройка метрик расстояния и фильтрации
- Оценка качества релевантности

### 3. Разработка ИИ агента на своей LLM модели

- Архитектура ИИ агента на основе LLM с контролем безопасности
- Техники снижения галлюцинаций

- Мультимодальные возможности (Llava, Bakllava)
- CJM создания ИИ-агента на кастомных данных

**Практическая часть:**

- Разработка специализированного ИИ агента на Llama 3
- Настройка system prompt для предметной области
- Интеграция с внешними API
- Тестирование и валидация ответов

#### **4. Разработка кастомного ИИ агента**

- Архитектура кастомных ИИ агентов
- Обзор opensource решений для разработки ИИ агентов
- Инструменты LangGraph, LangChain для создания агентов
- Графы состояний и многоагентные системы

**Практическая часть:**

- Реализация ИИ агента на основе LangGraph
- Интеграция с открытыми моделями reasoning (Qwen2.5)
- Создание multistep workflow
- Встраивание векторной БД как памяти агента

#### **5. Nocode платформы для быстрого прототипирования**

- использование No-Code/Low-code платформ для ИИ разработки (n8n, Langflow, LangGraph)
- Кейсы использования, когда выбирать No-Code vs код
- Интеграция с внешними API и базами данных

**Практическая часть:**

- Создание ИИ-агента в Langflow без программирования
- Интеграция с векторными базами данных через UI
- Развертывание прототипов в облако

#### **6. Развертывание и эксплуатация ИИ агентов в проде**

- Эксплуатация LLM в проде. фреймворки sglang/ollama
- MLOps для открытых LLM
- Развертывание, мониторинг, версионирование
- Мониторинг дрейфа и качества моделей

**Практическая часть:**

- Развертывание в продакшен с vLLM
- Настройка мониторинга с Prometheus/Grafana
- A/B тестирование моделей
- Оптимизация затрат на инференс