ИИ-инженер

01 Кому подойдёт курс

Backend-разработчикам:

Научитесь добавлять LLM в продукты: строить поиск по данным, разрабатывать AI-ассистентов и внедрять их в существующие сервисы

ML-инженерам:

Систематизируете работу с LLM: научитесь проектировать RAG-системы, управлять качеством генерации и выводить AI-решения в продакшен

DL / NLP-инженерам:

Расширите роль от моделирования к инженерии: будете создавать полноценные AI-системы с данными, логикой и интеграцией в бизнес-процессы

02 Чему научитесь на курсе

- Научитесь превращать LLM в работающие ИИ-продукты
- Настроите и оптимизируете LLM под бизнес-условия
- Построите RAG-систему с точным поиском и оценкой релевантности
- Разработаете AI-агентов с инструментами, памятью и контролем поведения
- Организуете оркестрацию ИИ-компонентов и асинхронную обработку запросов
- Выведете ИИ-сервис в продакшен с мониторингом и версионированием

03 Как проходит курс

- Сопровождение кураторами
- Обратная связь от опытных наставников
- Теория на платформе Практикума
- Практические задания с ревью на готовой инфраструктуре в облаке

Что вас ждёт на курсе

Диплом о профессиональной переподготовке или сертификат	Практика, основанная на решении реальных рабочих задач: 6 самостоятельных проектов в портфолио	Обучение от экспертов из Яндекса и других крупных компаний

ИИ-инженер

4 месяца
продолжительность курса

6 проектов
в портфолио

3 воркшопа
с наставником

3 НЕДЕЛИ | 36 ЧАСОВ

01

LLM в работе инженера: от архитектуры до управления генерацией

3 НЕДЕЛИ | 36 ЧАСОВ

02

RAG: от сырых данных до точных ответов

3 НЕДЕЛИ | 36 ЧАСОВ

03

Агентные системы: от одного агента до оркестрации

3 НЕДЕЛИ | 36 ЧАСОВ

04

Деплой ИИ-систем: от прототипа к работающему сервису

3 НЕДЕЛИ | 36 ЧАСОВ

05

Эксплуатация ИИ-систем: качество, безопасность и улучшение

2 НЕДЕЛИ | 24 ЧАСА

06

Итоговый проект. ИИ-ассистент для службы поддержки

LLM в работе инженера: от архитектуры до управления генерацией

01

3 недели | 36 часов
1 проект

Спринт даёт практическую основу для работы с языковыми моделями в продукте. Вы научитесь управлять качеством и стоимостью генерации, разберётесь с длинным контекстом как инженерной проблемой и освоите инструменты для эффективного инференса в реальных системах

Практическая работа
Соберёте LLM-сервис для конкретного бизнес-кейса: настроите параметры генерации, оцените качество и стоимость на реальных данных, подберёте оптимальные настройки модели под продакшен-сценарий

Содержание

01.	Как устроены современные языковые модели	• Токенизация, эмбеддинги и decoder-only архитектура • Как текст превращается в последовательность токенов, проходит через трансформер и используется для автогрессивной генерации следующего токена
02.	Управление генерацией и качеством ответов	• Temperature, top-p, top-k, repetition penalty и другие параметры генерации • Как они влияют на стиль ответов, устойчивость и воспроизводимость результатов в продукте
03.	Длинный контекст как продуктовая проблема	• Проблемы генерации на длинных последовательностях: деградация качества, рост latency и ошибки внимания. • Как разные подходы к позиционному кодированию (RoPE, ALiBi, YaRN) влияют на доступный контекст и выбор модели
04.	Инференс LLM в реальных системах	• Онлайн-инференс, batching и работа с KV-cache • Использование vLLM, flash-attention и квантизации для ускорения инференса и снижения стоимости

RAG: от сырых данных до точных ответов

02

3 недели | 36 часов
1 проект

Спринт посвящён построению систем поиска и генерации на основе RAG. Вы разберётесь, как подготовить данные, построить векторный поиск и соединить его с генерацией так, чтобы модель отвечала точно и по делу. Освоите полный пайплайн — от сырых документов до оценки качества системы на реальных данных

Практическая работа
Создадите рабочий RAG-пайплайн: подготовите данные, построите векторную базу, реализуете поиск и генерацию ответов с помощью LLM, проведёте оценку качества на реальных данных

Содержание

01.	Данные для RAG	• Очистка, нормализация, создание чанков и обработка документов разных форматов: pdf, txt, html • Как качество данных влияет на качество ответов
02.	Эмбеддинги и векторное представление	• Создание векторов с помощью трансформеров и OpenAI API • Применение снижения размерности через PQ • Понимание различий bi- и cross-encoder
03.	Векторные базы и индексы	• Проблемы генерации на длинных последовательностях: деградация качества, рост latency и ошибки внимания. • Как разные подходы к позиционному кодированию (RoPE, ALiBi, YaRN) влияют на доступный контекст и выбор модели
04.	Архитектура RAG	• Построение полного retrieval и генеративного пайплайна • Интеграция поиска и генерации, гибридные подходы для повышения качества и скорости ответов
05.	Оценка качества RAG	• Метрики retrieval и генерации, эксперименты с ранжированием • Использование LLM-as-judge для проверки релевантности и точности ответов системы

Агентные системы: от одного агента до оркестрации 03

3 недели | 36 часов
1 проект

Спринт посвящён созданию AI-агентов — систем, которые не просто отвечают, но и действуют. Вы научитесь строить агентов с инструментами, организовывать их взаимодействие и контролировать поведение. Разберётесь с мультиагентными архитектурами, оркестрацией и подключением внешних сервисов — включая работу с изображениями и веб-контентом

Практическая работа
Разработаете AI-агента, который выполняет цепочку задач с внешними инструментами и интернетом, оценивает свои ответы и интегрируется в рабочий сервис

Содержание

01.	Function calling и внешние инструменты	• Использование функций OpenAI и LangChain, structured output, настройка безопасных вызовов и обработка ошибок • Как дать модели возможность действовать через инструменты
02.	ReAct: строим агента и учим его рассуждать	• Построение ReAct-агента с набором инструментов • Как агент рассуждает, планирует и принимает решения • Использование LLM-as-judge для оценки качества и релевантности ответов агента
03.	AI guardrails и контроль поведения	• Фильтрация действий агента, предотвращение галлюцинаций, ограничение и проверка корректности поведения • Как сделать агента безопасным в продакшене
04.	Оркестрация LLM и пайплайны	• Chaining и пайплайны через LangChain и LangGraph • Управление последовательностью вызовов, разбиение задач и построение устойчивых сценариев работы агента
05.	Мультиагентные системы и интеграция внешних сервисов	• Создание мультиагентных систем: как агенты взаимодействуют и передают задачи друг другу • Интеграция с интернетом и внешними API: поиск, базы данных, внешние сервисы
06.	Мультимодальные модели и работа с изображениями	• Принцип работы Vision-Language моделей и практические задачи: анализ изображений, извлечение информации из документов, визуальный поиск • Как агент работает с нетекстовыми данными

Деплой ИИ-систем: от прототипа к работающему сервису 04

3 недели | 36 часов
1 проект

Спринт посвящён переходу от работающего прототипа к продакшен-сервису. Вы научитесь упаковывать LLM-приложение в контейнер, строить API, настраивать асинхронную обработку запросов и мониторить систему — с фокусом на метриках, которые важны именно для LLM: latency, стоимость токенов, качество ответов

Практическая работа
Соберёте продакшен-сервис ИИ: упакуете LLM, RAG и агента в единую систему с обработкой реальных запросов

Содержание

01.	Деплой LLM-сервиса	• Упаковка LLM-приложения в Docker-контейнер, создание FastAPI-сервиса, настройка параметров инференса • Фокус не на Docker как инструменте — а на том, как быстро и надёжно выкатить LLM-сервис в продакшен
02.	Асинхронность и батчинг	• Ускорение обработки запросов, управление нагрузкой и latency, оптимизация потоков данных под реальную нагрузку • Как LLM-сервис справляется с несколькими запросами одновременно
03.	Мониторинг LLM-сервиса: метрики и observability	• Что измерять в LLM-системе: latency, стоимость токенов, throughput, процент ошибок • Трекинг цепочек вызовов и отладка через LangSmith • Базовый дашборд в Grafana для AI-системы
04.	Интеграция: LLM + RAG + агент	• Сборка полного пайплайна из компонентов, вызовы API, обработка ошибок, контроль корректности и качества на стыке систем

Эксплуатация ИИ-систем: качество, безопасность и улучшение

05

3 недели | 36 часов
1 проект

Спринт посвящён тому, что происходит после запуска. Вы научитесь защищать систему от атак и непредсказуемого поведения, измерять качество на всех уровнях — генерации, RAG и агента — и принимать осознанные решения об улучшении: от настройки параметров до дообучения модели

Практическая работа
Возьмёте продакшен-сервис, оцените качество системы на реальных данных и улучшите её — через настройку параметров или дообучение модели

Содержание

01.	Безопасность и надёжность LLM-систем	• Prompt injection и джейлбрейк, утечки данных через контекст, контроль рисков • Как сделать AI-систему устойчивой к атакам и непредсказуемому поведению
02.	Оценка качества AI-системы	• Метрики качества генерации, RAG и агента. LLM-as-judge, автоматические метрики и подходы к оценке на реальных данных • Как понять, что система работает хорошо — и где она проваливается
03.	Когда и как дообучать модель	• Обзор подходов к адаптации LLM: LoRA, QLoRA и их ограничения • В каких задачах дообучение оправдано — а когда достаточно правильно настроить инференс или промпт • Практика: запускаем LoRA на реальной задаче
04.	Улучшение и оптимизация системы	• Как находить узкие места в работающей системе и устранять их • Оптимизация производительности и качества под реальные бизнес-ограничения

Итоговый проект. ИИ-ассистент для службы поддержки

06

2 недели | 24 часа

Вы создадите ИИ-ассистента для службы поддержки компании: спроектируете архитектуру, подготовите данные, соберёте RAG-пайплайн, реализуете агента с инструментами и выкатите систему в продакшен с мониторингом и оценкой качества

Проект объединяет всё, что пройдено на курсе — от первого запроса к модели до работающего сервиса

Итоговый проект. ИИ-ассистент для службы поддержки

06

2 недели | 24 часа

Вы создадите ИИ-ассистента для службы поддержки компании: спроектируете архитектуру, подготовите данные, соберёте RAG-пайплайн, реализуете агента с инструментами и выкатите систему в продакшен с мониторингом и оценкой качества

Проект объединяет всё, что пройдено на курсе — от первого запроса к модели до работающего сервиса

Медведев Кирилл
Директор АНО ДПО «Образовательные технологии Яндекса»