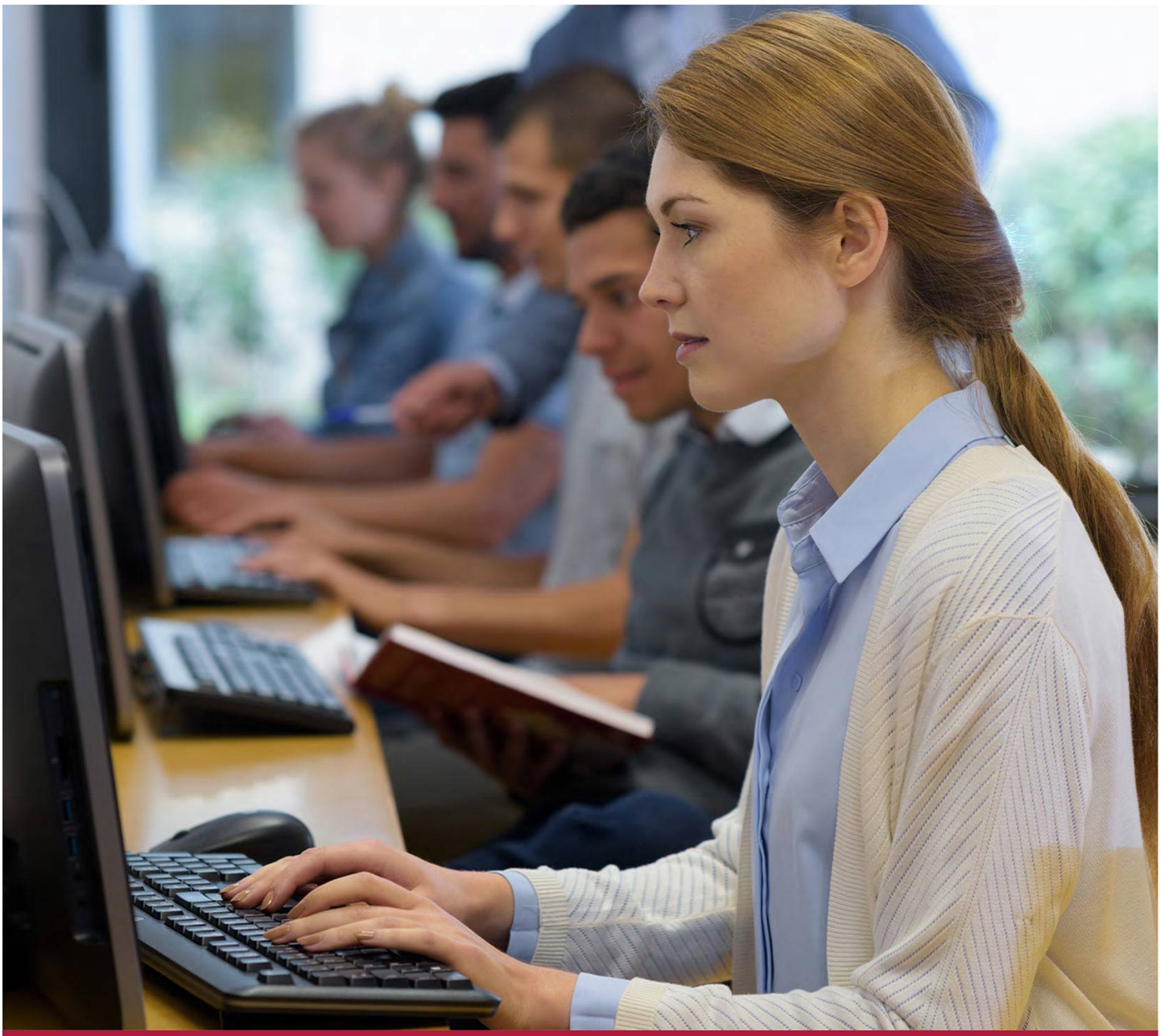
Академия АйТи
a Softline Company



LangChain Enterprise: От локальной модели до агентных систем

Код курса: LANGCHAIN

+7 (495) 150 96 00 | academy@academyit.ru | academyit.ru

LangChain Enterprise: От локальной модели до агентных систем

Код курса: LANGCHAIN

Длительность	8 ак. часов
Формат	
Разработчик курса	Академия АйТи
Тип	Учебный курс
Способ обучения	Под руководством тренера

О курсе

8-часовой мастер-класс о создании on-premise стека и приложений на базе LLM. За день участники разворачивают инфраструктуру через Docker, знакомятся с локальными моделями (в т.ч. Qwen-2.5) и фреймворками запуска (Ollama, vLLM, SGLang), подключают LLM в LangChain, используют LCEL и структурированный вывод с Pydantic, а также работают с мультимодальной LLaVA для анализа изображений. Далее на практике строится RAG над корпоративными документами с векторными БД, полнотекстовым и гибридным поиском. Рассматривается извлечение троек и построение графов знаний, проектирование агентных workflow в LangGraph, стратегии работы с ограниченным контекстом и основы вывода в продакшен: ключевые метрики, логирование/трассировка и простой UI на Gradio.

Подробная информация

Профиль аудитории:

- разработчики и инженеры, вовлечённые во внедрение ИИ/LLM: ML/DS-инженеры, backend-разработчики, инженеры по данным и архитекторы решений; аналитики, которым нужно собирать прототипы ассистентов и RAG-систем на собственных данных

Предварительные требования:

- базовые навыки Python; умение работать с Docker/docker-compose и командной строкой; понимание принципов работы веб-API (CLI/API вызовы)

По окончании курса слушатели смогут:

- разворачивать локальный стек LLM через Docker/docker-compose и запускать модели (например, Qwen, Llama) через CLI и API
- подключать локальную модель к приложению через LangChain и получать структурированный вывод (Pydantic/JSON)
- собирать базовый RAG над документами: загрузка/чанкинг, индексация в ChromaDB, полнотекстовый и гибридный поиск и включение результатов в цепочку
- извлекать факты и строить простой граф знаний с поиском по нему

- проектировать и реализовывать агента в LangGraph: состояние/узлы/рёбра, Tool Calling, условный роутинг и базовый роутер инструментов
- учитывать ограничение контекста: применять сжатие истории, Sliding Window, суммаризацию и простой fallback при ошибках/зацикливаниях
- поднимать простой веб-интерфейс на Gradio и включать базовое логирование шагов исполнения

Программа курса

Модуль 1. Локальная инфраструктура

- Почему on-premise: безопасность данных, контроль затрат, кастомизация, независимость
- Обзор локальных LLM: сравнение моделей; Qwen-2.5 как оптимальный выбор для enterprise
- Сравнение фреймворков запуска: Ollama (прототипы), vLLM (production-пропускная способность), SGLang (скорость для сложных промптов и RAG)
- Развёртывание стека через docker-compose
- Загрузка и тестовые прогоны моделей (Qwen, Llama) через CLI и API

Модуль 2. Основы LangChain + локальные модели

- Архитектура LangChain для on-prem: перенаправление на локальные эндпоинты
- LCEL: декларативные цепочки со streaming, async и параллелизмом
- Структурированный вывод с Pydantic (валидный JSON)
- Подключение к локальной LLM через ChatOllama
- Создание первой LCEL-цепочки и парсинг вывода в JSON/Pydantic
- Мультимодальная LLaVA: анализ изображений

Модуль 3. Поисковые системы и RAG

- Введение в Retrieval-Augmented Generation (RAG)
- Векторный поиск: от ChromaDB к enterprise-решениям (Qdrant, Milvus)
- Полнотекстовый поиск (BM25) и гибридный поиск
- Загрузка и чанкинг корпоративных документов
- Создание векторной базы в ChromaDB
- Сборка RAG-цепочки на LCEL
- Реализация гибридного поиска с EnsembleRetriever

Модуль 4. Графы знаний

- Что такое графы знаний и как они дополняют RAG
- Извлечение троек (субъект-предикат-объект) с помощью LLM
- Создание цепочки для извлечения троек
- Построение и визуализация графа (NetworkX)
- Кастомный ретривер для поиска по графу

Модуль 5. Агентные системы и LangGraph

- Отличие агентов от цепочек: динамическое принятие решений
- Архитектура LangGraph: State, Nodes, Edges

- Tool Calling для вызова внешних функций
- Создание кастомных инструментов
- Построение графа состояний и условный роутинг
- Стратегия для 20+ инструментов: агент-роутер (категоризация → выбор инструмента)

Модуль 6. Продвинутые кейсы и ограничения

- Ограниченное контекстное окно: проблемы и подходы
- Стратегии: сжатие истории, Sliding Window, явная суммаризация
- Реализация сжатия истории в стейте LangGraph
- Применение Sliding Window для простых случаев
- Fallback-механизм на счётчике рекурсии (зацикливания и ошибки агента)

Модуль 7. Мониторинг и production

- Ключевые метрики: Latency, Accuracy, Token Usage
- Логирование и трассировка вызовов
- Создание простого UI с Gradio
- Callbacks для логирования шагов агента в консоль (прототип для ELK/Datadog)

Посмотреть расписание курса и записаться на обучение

Обращайтесь по любым вопросам
к менеджерам Академии АйТи

+7 (495) 150 96 00 | academy@academyit.ru