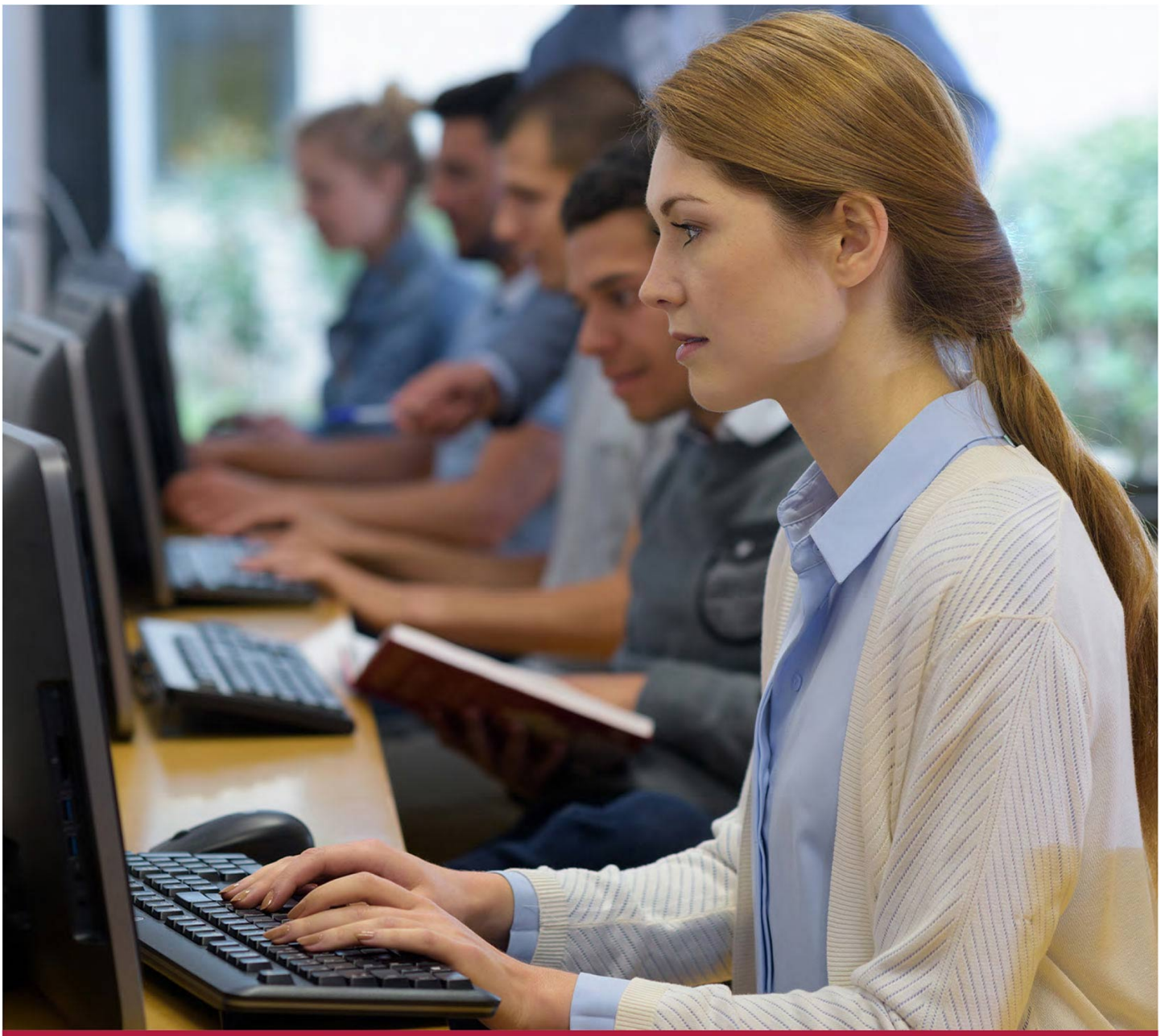
Академия АйТи
a Softline Company



Архитектура современных RAG-систем: от векторов до GraphRAG и агентов

Код курса: RAG-1

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

Архитектура современных RAG-систем: от векторов до GraphRAG и агентов

Код курса: RAG-1

Длительность	40 ак. часов
Формат	
Разработчик курса	Академия АйТи
Тип	Учебный курс
Способ обучения	Под руководством тренера

О курсе

Комплексный курс по проектированию и внедрению отказоустойчивых систем генерации с извлечением знаний (RAG). Программа охватывает весь спектр современных подходов: от базового Naive RAG через продвинутые методы поиска и GraphRAG до адаптивных (Self-RAG, CRAG) и агентных архитектур. Курс включает работу с мультимодальными данными, оценку качества и подготовку систем к production.

Подробная информация

Профиль аудитории:

- ML-инженеры и NLP-специалисты, разрабатывающие системы на основе LLM
- Архитекторы AI-решений, проектирующие RAG-пайплайны
- Backend-разработчики, интегрирующие RAG-системы в продукты
- Data Engineers, работающие с векторными базами данных и графами знаний

Предварительные требования:

- Опыт работы с Python и LLM-фреймворками (LangChain, LlamaIndex)
- Базовое понимание архитектуры Transformer, эмбеддингов и векторного поиска
- Знакомство с Docker и REST API
- Опыт работы с базами данных (SQL или NoSQL)

По окончании курса слушатели смогут:

- Проектировать и реализовывать RAG-пайплайны от Naive RAG до продвинутых архитектур
- Применять продвинутые методы поиска: гибридный поиск (BM25 + векторный), Query Expansion, HyDE, Re-ranking
- Строить GraphRAG-системы на базе Neo4j с извлечением сущностей и графовым поиском
- Реализовывать адаптивные подходы: Corrective RAG, Self-RAG, LongRAG
- Проектировать Agentic RAG с многошаговым рассуждением и семантической маршрутизацией
- Работать с мультимодальными данными (изображения, таблицы, графики) в RAG
- Оценивать качество RAG-систем (RAGAS, DeepEval) и обеспечивать масштабирование и

безопасность

Программа курса

Модуль 1. Фундамент: Naive RAG и оптимизация данных

- Тема 1.1. Архитектура Naive RAG
- Ограничения контекстного окна и проблема галлюцинаций. Компоненты: ETL-пайплайн, Embedding-модели, Vector Stores.
- Тема 1.2. Продвинутый Chunking (Разбиение)
- От фиксированного окна к семантическому разбиению. Рекурсивное разбиение и учет разметки (Markdown/HTML/PDF).
- Тема 1.3. Векторные базы данных 2026
- Сравнение: Pinecone, Weaviate, Milvus, Chroma, pgvector. Индексация: HNSW против IVFFlat.
- Практика: Создание базового RAG-пайплайна на LlamaIndex с семантическим чанкингом.

Модуль 2. Advanced Retrieval: Качество поиска

- Тема 2.1. Гибридный поиск (Hybrid Search)
- Соединение векторного поиска и BM25. Reciprocal Rank Fusion (RRF).
- Тема 2.2. Продвинутая обработка запросов
- Query Expansion, Multi-Query, Query Translation. Гипотетические эмбеддинги (HyDE).
- Тема 2.3. Re-ranking (Переранжирование)
- Cross-Encoders, Cohere Rerank, BGE-Reranker.
- Практика: Внедрение переранжирования и оценка прироста Hit Rate.

Модуль 3. GraphRAG: Интеллект на графах знаний

- Тема 3.1. Введение в Knowledge Graphs
- Почему векторы не видят связей? Графы как способ хранения онтологий. Инструменты: Neo4j, FalkorDB.
- Тема 3.2. Архитектура Microsoft GraphRAG
- Извлечение сущностей и связей. Генерация коммьюнити-репортов (Community Summaries).
- Тема 3.3. Глобальный против Локального поиска
- Когда использовать векторный поиск, а когда — графовый.
- Практика: Развертывание GraphRAG на Neo4j и сравнение с классическим RAG.

Модуль 4. Адаптивные и итеративные подходы

- Тема 4.1. Corrective RAG (CRAG)
- Автоматическая оценка релевантности и запуск веб-поиска при нехватке знаний.
- Тема 4.2. Self-RAG
- Обучение модели критиковать собственные ответы. Специальные токены рефлексии.
- Тема 4.3. LongRAG
- Стратегии для огромных контекстных окон (1M+ токенов). RAG vs. длинный контекст.
- Практика: Реализация самопроверки (Self-Correction) в пайплайне на LangGraph.

Модуль 5. Agentic RAG: Агенты-исследователи

- Тема 5.1. RAG как инструмент Агента
- Паттерн ReAct. Агентный выбор между поиском в документах, SQL-базе или интернете.
- Тема 5.2. Многошаговое рассуждение (Multi-hop Reasoning)
- Декомпозиция сложного вопроса на подвопросы. Сбор информации из разных источников.
- Тема 5.3. Маршрутизация (Semantic Routing)
- Динамический выбор лучшего пайплайна в зависимости от интента пользователя.
- Практика: AI-агент для техподдержки — поиск в базе знаний или запрос статуса заказа через API.

Модуль 6. Мультимодальный RAG и Production

- Тема 6.1. Multi-modal RAG
- Работа с изображениями, таблицами и графиками. Визуальные эмбеддинги (CLIP, ColPali).
- Тема 6.2. Оценка систем (RAG Evaluation)
- Фреймворки RAGAS и DeepEval. Метрики: Faithfulness, Answer Relevance, Context Precision.
- Тема 6.3. Масштабирование и Безопасность
- Semantic Caching. Защита от Prompt Injection через контекст. Оптимизация стоимости токенов.
- Практика: Автоматическое тестирование RAG-системы с помощью RAGAS и генерация отчета.

Посмотреть расписание курса и записаться на обучение

Обращайтесь по любым вопросам
к менеджерам Академии АйТи

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