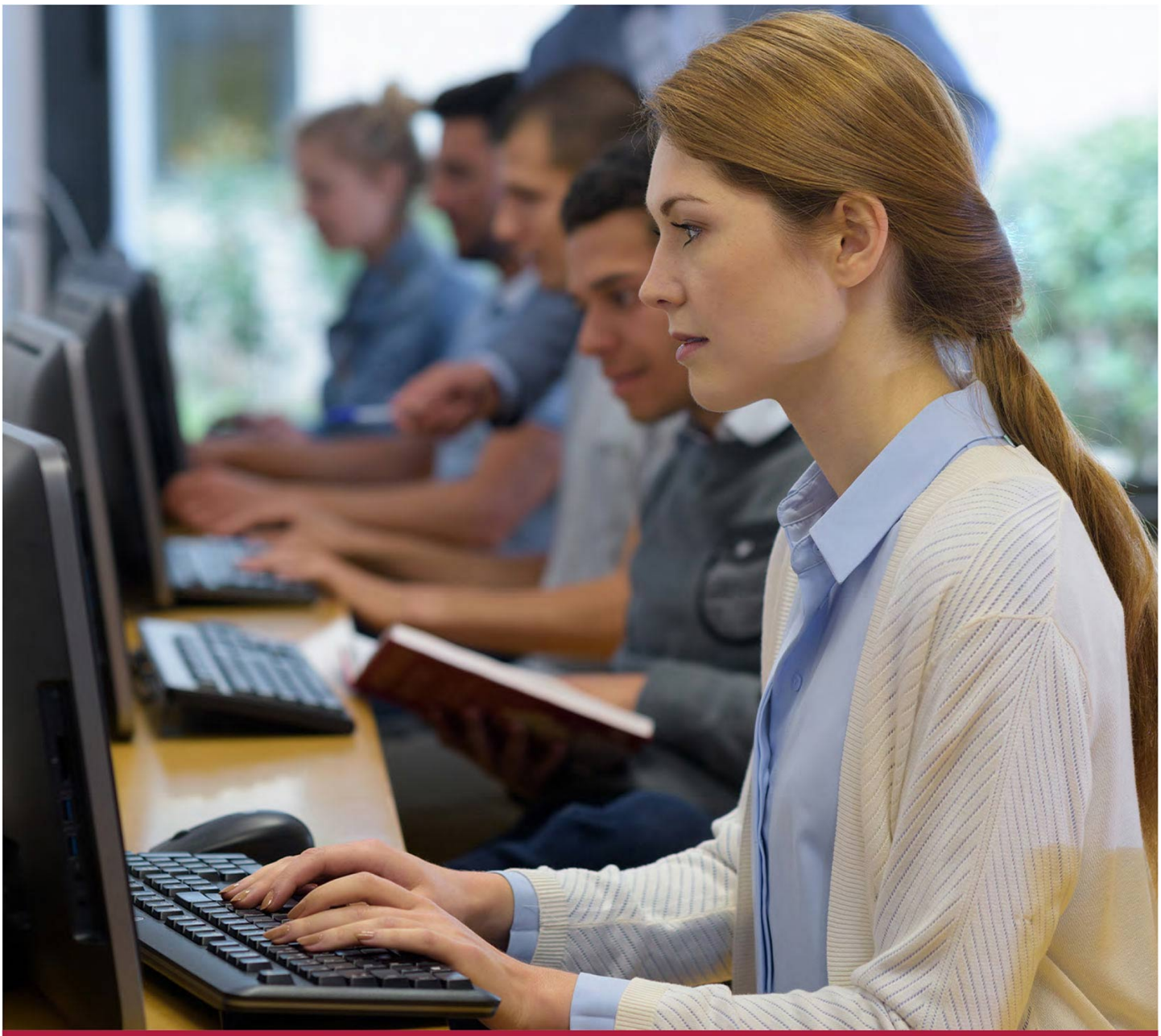
Академия АйТи
a Softline Company



Архитектура генеративных ИИ-моделей

Код курса: ARCH-GENAI

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

Архитектура генеративных ИИ-моделей

Код курса: ARCH-GENAI

Длительность	72 ак. часа
Формат	
Разработчик курса	Академия АйТи
Тип	Учебный курс
Способ обучения	Под руководством тренера

О курсе

Комплексный курс для руководителей и экспертов, формирующий системное видение роли архитектора в проектировании и внедрении GenAI-решений. Программа охватывает полный жизненный цикл проекта: от пресейла и работы с требованиями до развертывания высоконагруженных AI-систем в production. Особое внимание уделяется on-premise решениям, Open Source инструментам и Low-code автоматизации в контексте интегрированного управления рисками.

Подробная информация

Профиль аудитории:

- Руководители служб и начальники отделов в области управления рисками, комплаенса и внутреннего контроля
- Архитекторы программного обеспечения и системные аналитики
- Эксперты, планирующие внедрение AI-решений в корпоративные бизнес-процессы
- Технические лидеры, отвечающие за стратегию цифровой трансформации

Предварительные требования:

- Опыт работы в IT или в области управления рисками от 3 лет
- Базовое понимание архитектуры программного обеспечения и жизненного цикла разработки
- Общее представление о технологиях искусственного интеллекта и машинного обучения
- Опыт работы с документацией (ТЗ, SRS, ADR) будет преимуществом

По окончании курса слушатели смогут:

- Анализировать проектные ограничения, выявлять риски и планировать AI-проект как последовательность этапов, приносящих измеримую ценность
- Проектировать и документировать архитектуру AI-решений с использованием C4 Model, ADR и OpenAPI-спецификаций
- Применять архитектурные паттерны RAG, AI-агентов и мультиагентных систем
- Встраивать механизмы обеспечения качества, безопасности и наблюдаемости в архитектуру AI-систем

- Рассчитывать ресурсы для инференса LLM, проектировать IaC и CI/CD пайплайны
- Применять продвинутые архитектурные паттерны: Serverless, EDA, High-Load, Multi-tenancy, Federated Learning
- Управлять стоимостью облачных решений (FinOps), выстраивать этическую AI-архитектуру и коммуницировать с C-Level

Программа курса

Модуль 1. Стратегический фундамент и планирование проекта

- Цель: Сформировать системное видение роли архитектора в бизнес-процессах. Научить анализировать проектные ограничения, выявлять риски и планировать проект как последовательность этапов, поставляющих измеримую ценность.
- Пресейл, контракты и работа с требованиями: закладываем фундамент проекта
- Теория: Роль архитектора на этапе пресейла. Анализ исходных требований. Типология проектов и модели контрактов (FP, T&M, T&M with a cap).
- Практика: Кейс-стади — анализ трех обезличенных запросов (RFP/ТЗ), выбор и обоснование модели контракта.
- Проектирование и оценка: от требований к плану, рискам и смете
- Теория: Жизненный цикл (SDLC) и формализация требований (SRS). Техники оценки проектов (Analogous, Parametric, PERT, Bottom-Up). Управление рисками в Fixed Price проектах.
- Практика: Декомпозиция User Story, оценка задачи, разработка матрицы рисков и черновика Change Request.
- Стратегия поставки ценности: от PoC до Production
- Теория: Этапы поставки ценности (Демо, PoC, MVP). Связь стратегии с контрактом. Путь в Production.
- Практика: Разработка дорожной карты (Roadmap) для проекта с AI-рекомендациями.

Модуль 2. Проектирование и документирование архитектуры

- Цель: Дать полный набор инструментов для создания, документирования и верификации архитектуры AI-решений.
- Высокоуровневое проектирование (HLD) с использованием C4 Model
- Теория: Методология C4. Диаграммы Уровня 1 (Контекст) и Уровня 2 (Контейнеры). Практика: Создание диаграмм C1 и C2.
- Низкоуровневое проектирование (LLD): компоненты и взаимодействия
- Теория: Детализация HLD. Диаграммы Уровня 3, Sequence Diagrams, OpenAPI. Практика: Разработка LLD для одного контейнера.
- Архитектурные паттерны: RAG и его продвинутые вариации
- Теория: Пайплайны RAG, Self-RAG, CRAG, Knowledge/Cache Augmented Generation. Практика: Проектирование гибридной RAG-архитектуры.
- Архитектурные паттерны: AI-агенты и Multi-Agent Systems
- Теория: Агентные циклы (ReAct, Plan-and-Execute). Паттерны мультиагентных систем. Практика: Проектирование мультиагентной системы.
- Документирование решений: Architecture Decision Records (ADR)
- Теория: Формат и назначение ADR. AAC — архитектура как код. Практика: Написание ADR.
- Верификация архитектуры и CTO Challenge
- Теория: Методы верификации. CTO Challenge как практика аудита. Практика: Ролевая игра —

сессия архитектурного ревью.

Модуль 3. Качество, интеграции и безопасность

- Цель: Научить встраивать в архитектуру механизмы обеспечения качества, надежности и безопасности.
- Архитектурный надзор и управление техническим долгом
- Теория: Архитектурный надзор, код-ревью, технический долг. Практика: Ревью pull request, документирование технического долга.
- Проектирование интеграций: от классики до AI-стандартов
- Теория: Протоколы (HTTP, SMTP, gRPC), паттерны асинхронной интеграции, протоколы A2A и MCP. Практика: Проектирование отказоустойчивой интеграции.
- Архитектура данных для AI-систем
- Теория: Data pipelines, Data Lake/Warehouse/Lakehouse, Feature Store, Data Governance. Практика: Проектирование end-to-end data pipeline.
- Оценка качества и тестирование GenAI-компонентов
- Теория: Метрики для RAG (Faithfulness, Answer Relevancy), DeepEval, Ragas. Практика: Разработка плана тестирования для RAG-задачи.
- Security by Design: архитектура для защиты AI-систем
- Теория: Security by Design, OWASP Top 10 для LLM, Guardrails, PII Sanitization. Практика: Усиление архитектуры AI-агента компонентами безопасности.
- Архитектура наблюдаемости (Observability)
- Теория: Метрики, Логи, Трейсы. Prometheus, Grafana, Jaeger. Практика: Проектирование дашборда для мониторинга SLO AI-сервиса.

Модуль 4. Инфраструктура

- Цель: Дать системные знания по планированию, автоматизации развертывания и поддержке инфраструктуры для AI-систем.
- Расчет ресурсов (Sizing) для приложений и данных
- Теория: Методология расчета ресурсов (CPU, RAM, Disk). Влияние RPS и финансовых ограничений. TCO для on-premise. Практика: Расчет ресурсов для учебного проекта.
- Расчет ресурсов и оптимизация инференса LLM
- Теория: Расчет VRAM, квантизация, FlashAttention, vLLM, непрерывный батчинг. Практика: Выбор оптимальной конфигурации GPU-инстанса.
- Инфраструктура как код (IaC) и CI/CD
- Теория: Принципы IaC (Terraform), CI/CD пайплайн с quality gates. Практика: Terraform-конфигурация для базовой инфраструктуры.
- Архитектура MLOps-конвейеров
- Теория: Конвейеры для дообучения моделей (Airflow, Kubeflow), мониторинг model drift. Практика: Проектирование MLOps-пайплайна для RAG-системы.
- Стратегии развертывания и вывода в Production
- Теория: Blue-Green, Canary, Rolling. Чек-лист готовности, план отката. Практика: План релиза с Canary-стратегией.
- Архитектура высокой доступности (HA) и восстановления (DR)
- Теория: Паттерны отказоустойчивости, RTO/RPO, гео-резервирование. Практика: HA-архитектура для критически важного AI-сервиса.

Модуль 5. Продвинутые архитектурные паттерны

- Цель: Изучить передовые архитектурные подходы для масштабирования, real-time обработки и работы в гибридных средах.
- Serverless vs. Kubernetes для AI-ворклоадов
- Теория: Сравнительный анализ, трейдоффы. Практика: Архитектурный воркшоп для набора сценариев.
- Событийно-ориентированная архитектура (EDA) для AI
- Теория: Принципы EDA, паттерны Pub/Sub, Event Sourcing, CQRS. Практика: EDA-архитектура для fraud detection.
- Архитектура для High-Load и Low-Latency инференса
- Теория: Кэширование, Edge computing для AI. Практика: Архитектура рекомендательного сервиса с latency < 50 мс.
- Гибридная и мультиоблачная архитектура для AI
- Теория: Cloud Bursting, Anthos, OpenShift. Практика: Гибридная архитектура — обучение on-premise, инференс в облаке.
- Архитектура для Multi-tenancy в AI SaaS
- Теория: Паттерны изоляции (Silo, Pool, Bridge). Практика: Архитектура безопасности для multi-tenant RAG-сервиса.
- Federated Learning и Privacy-Preserving архитектура
- Теория: Federated Learning, Differential Privacy. Практика: Архитектура федеративного обучения для медицинских данных.

Модуль 6. Стратегия, лидерство и экономика

- Цель: Развить стратегическое мышление, экономическую ответственность и лидерские качества.
- FinOps: архитектура, управляемая стоимостью
- Теория: Принципы FinOps. Практика: Анализ cloud bill, оптимизация затрат на 20%.
- Технологический радар и эволюция архитектуры
- Теория: Технологический радар, эволюционная архитектура. Практика: Сессия наполнения техрадара.
- Ethical AI by Design и архитектура для Governance
- Теория: Ответственный AI. Практика: Проектирование Model Card и архитектуры для аудита AI.
- API как продукт: проектирование и управление
- Теория: Проектирование API, версионирование, монетизация. Практика: Стратегия развития AI-сервиса в API-продукт.
- Техническое лидерство и коммуникация с C-Level
- Теория: Техническое лидерство, менторинг, презентация решений. Практика: Питч "Зачем инвестировать в EDA".
- Финальная защита комплексного проекта
- Практика: Студенты представляют итоговые проекты в формате презентации для архитектурного комитета.

Посмотреть расписание курса и записаться на обучение

Обращайтесь по любым вопросам
к менеджерам Академии АйТи

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