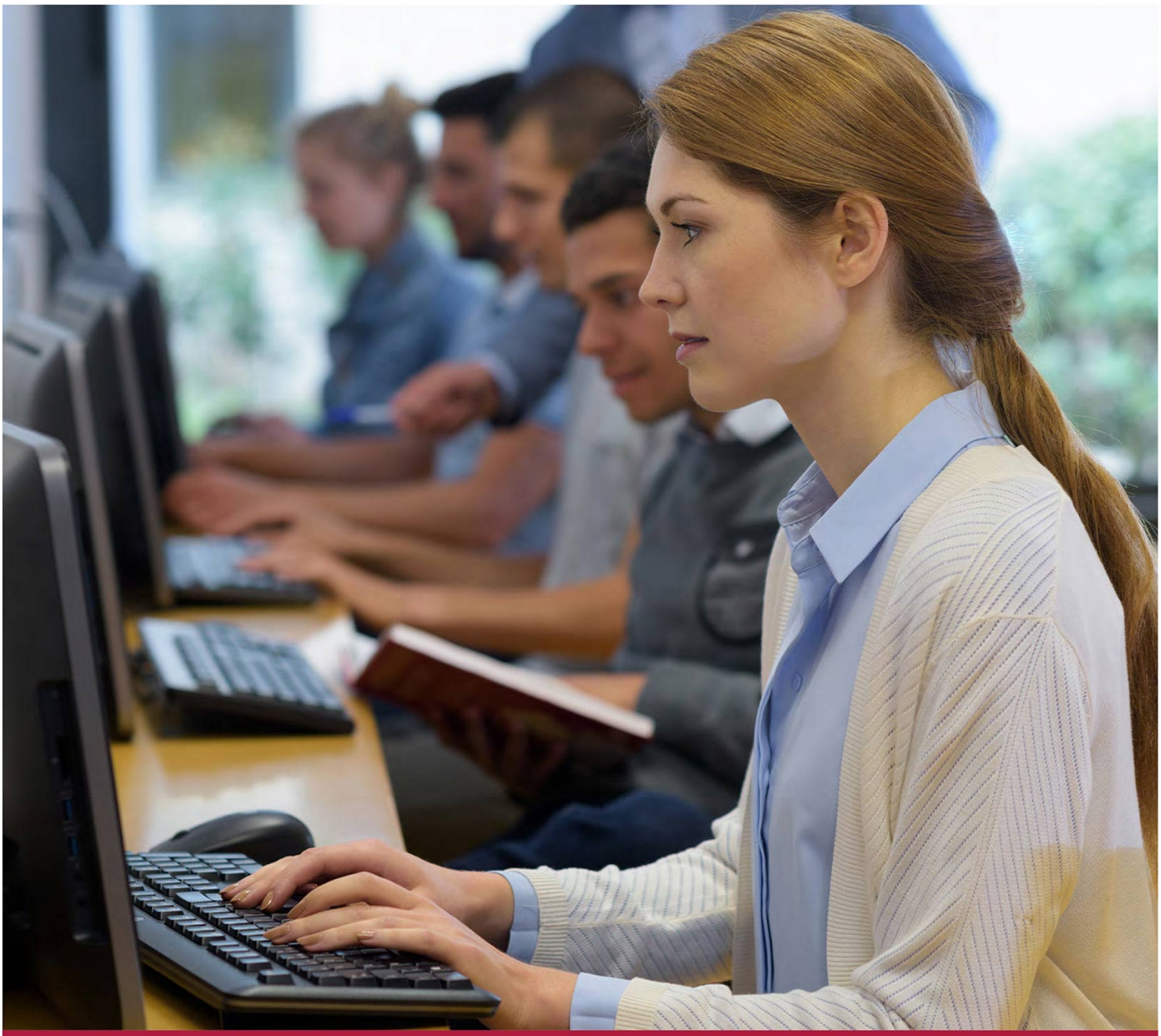
Академия АйТи
a Softline Company



Контролируемая генерация и безопасность LLM-систем

Код курса: LLM-SEC

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

Контролируемая генерация и безопасность LLM-систем

Код курса: LLM-SEC

Длительность	36 ак. часов
Формат	
Разработчик курса	Академия АйТи
Тип	Учебный курс
Способ обучения	Под руководством тренера

О курсе

Курс для инженеров и разработчиков, обучающий управлению поведением нейросетей, минимизации галлюцинаций, внедрению систем защиты от атак и настройке мониторинга качества в реальном времени. Программа охватывает весь спектр методов контроля: от промпт-инжиниринга и RAG до манипуляции логитами, Guardrails, промышленного мониторинга и защиты от adversarial-атак по стандартам OWASP Top 10 для LLM.

Подробная информация

Профиль аудитории:

- Инженеры и разработчики, работающие с большими языковыми моделями
- ML-инженеры, отвечающие за качество и безопасность AI-систем
- DevOps/MLOps-специалисты, внедряющие LLM в production
- Архитекторы AI-решений, проектирующие системы с контролем генерации

Предварительные требования:

- Опыт работы с LLM (GPT, LLaMA, Qwen)
- Знание Python и основных ML-фреймворков (PyTorch, Hugging Face)
- Базовое понимание архитектуры Transformer и механизма attention
- Опыт работы с LangChain или аналогичными фреймворками будет преимуществом

По окончании курса слушатели смогут:

- Применять техники промпт-инжиниринга (Zero-shot, Few-shot) для получения детерминированных ответов
- Строить RAG-системы и использовать графы знаний (KGA) для расширения знаний модели
- Применять техники рассуждения (CoT, ToT, GoT) и проектировать мультиагентные системы
- Управлять генерацией на уровне логитов: Top-K, Top-P фильтрация, Allowlist/Blocklist
- Настраивать параметры генерации (Temperature) под конкретные задачи
- Внедрять системы Guardrails для валидации входов и выходов модели
- Настраивать промышленный мониторинг (Prometheus, Grafana) и защищать LLM от атак

(Prompt Injection, Jailbreak)

Программа курса

Блок 1. Основы управления генерацией (Prompt Engineering)

- Урок 1. Введение: природа галлюцинаций и архитектура контроля
- Теория: Проблема отсутствия критического мышления у LLM. Эволюция от RNN/LSTM к Transformer и Attention. Техники: Zero-shot, One-shot, Few-shot, системный промпт. Практика: Чат-бот для магазина на LangChain.
- Урок 2. Расширение знаний: RAG и графы знаний (KGA)
- Теория: RAG vs. KGA. Методы поиска: векторный (Chroma/Weaviate), BM25, гибридный. Практика: RAG-система на LlamaIndex для ответов по статьям из Википедии.

Блок 2. Логика и агентные архитектуры

- Урок 3. Техники рассуждения (CoT, ToT, GoT)
- Теория: Chain-of-Thought, Tree-of-Thought, Graph-of-Thought. Агенты: Zero-shot ReAct, Plan-and-execute. Практика: Кастомный агент для решения логических головоломок.
- Урок 4. Мультиагентные системы и использование инструментов
- Теория: Архитектура "оркестра" агентов. Специализация ролей. Интеграция API. Практика: Система планирования путешествий из трех агентов.

Блок 3. Технический контроль вывода и тюнинг

- Урок 5. Манипуляция логитами и постобработка
- Теория: Работа с вероятностями токенов, Allowlist/Blocklist. Top-K vs. Nucleus Sampling (Top-P). Практика: Фильтр токсичности на TinyLlama.
- Урок 6. Тонкая настройка параметров (Parameter Tuning)
- Теория: Temperature, баланс креативности и точности. Шпаргалка по задачам. Практика: Исследование влияния настроек на Qwen 2.5.

Блок 4. Безопасность, Guardrails и эксплуатация

- Урок 7. Guardrails: система "защитных поручней"
- Теория: Этические и корпоративные ограничения. JSON Schema валидация. Многоэтапные пайплайны. Практика: Система apply_guardrails для Llama-2.
- Урок 8. Промышленный мониторинг и отладка
- Теория: Метрики качества (Perplexity, BLEU/ROUGE), безопасности (Toxicity), производительности. Стек Prometheus + Grafana + Jaeger + Loki. Паттерны Circuit Breaker и Health Check. Практика: Диагностика "бутылочных горлышек".
- Урок 9. Защита от атак (Adversarial Attacks & Defense)
- Теория: OWASP Top 10 для LLM 2025. Prompt Injection, Jailbreak, Data Poisoning, Supply Chain Vulnerabilities. Защита: RA-LLM, Adversarial Training, Differential Privacy. Практика: Red-teaming и многоуровневая фильтрация запросов.

Посмотреть расписание курса и записаться на обучение

Обращайтесь по любым вопросам
к менеджерам Академии АйТи

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