ФАСТ ЛЕЙН
УЧЕБНЫЙ ЦЕНТР

Безопасность генеративного ИИ — от атак до устойчивой защиты генеративных и агентных приложений (GENAI_SECURITY)

ID OT-GENAI_SECURITY **Цена** 145 000,– руб. **Длительность** 5 дней

Кому следует посетить

- Руководители и менеджеры по информационной безопасности (ИБ)
- Инженеры по безопасности (специалисты Blue Team и Red Team)
- Специалисты MLOps и ML-инженеры
- Разработчики LLM-агентов и чат-ботов
- Архитекторы сервисов на базе искусственного интеллекта

Предварительные требования

Для успешного прохождения курса рекомендуется базовое понимание принципов машинного обучения и информационной безопасности. Опыт практической работы с ИИ-моделями или в сфере кибербезопасности будет полезен, однако курс рассчитан и на слушателей без глубокого погружения в эти области.

Цели курса

После окончания курса участники смогут:

- Оценивать риски безопасности в проектах с генеративным ИИ и агентными системами, распознавая основные угрозы (от промпт-инъекций до адверсариальных атак).
- Проводить Red Team-тестирование моделей (LLM) для выявления уязвимостей: обнаруживать джейлбрейки, некорректное поведение моделей и другие способы обхода ограничений.
- Реализовывать меры Blue Team-защиты, чтобы повысить устойчивость ИИ-систем к атакам (например, защиту системных промптов, контроль доступа к моделям, мониторинг).
- Учитывать актуальные правовые нормы и стандарты (например, принципы доверенного ИИ, требования AI Act, стандарт ISO/IEC 42001) при разработке и внедрении ИИ-сервисов.
- Применять методологии моделирования угроз для ИИ (MITRE, NIST, MAESTRO и др.) и использовать специализированные инструменты мониторинга и тестирования на практике.
- Внедрять процессы непрерывного тестирования и аудита безопасности ИИ-моделей, своевременно выявляя и устраняя новые уязвимости по мере их появления.

Содержание курса

Генеративные модели искусственного интеллекта открывают бизнесу новые возможности, но одновременно создают и уникальные риски. «Безопасность генеративного ИИ» – это курс, который охватывает весь спектр этих новых вызовов: от понимания того, как злоумышленники могут атаковать ИИ-системы, до освоения эффективных мер защиты и обеспечения соответствия регуляторным требованиям. Программа строится вокруг практических сценариев: участники изучат реальные инциденты 2024–2025 годов, разберут механизмы промпт-инъекций и адверсариальных атак, а затем научатся противодействовать им с помощью современных инструментов и методик.

Обучение включает 8 модулей (лекции с демонстрациями, ~4 академических часа каждый) и 3 практических проекта **в формате Red Team / Blue Team**. В ходе этих проектов участники на практике попробуют взломать условного ИИ-агента и разработать меры защиты, закрепляя полученные знания. Такой баланс теории и практики позволит приобрести не только знания, но и ценные навыки: все ключевые методики отрабатываются на примерах. К концу курса слушатели будут чётко понимать, что может пойти не так при внедрении генеративного ИИ и какие шаги нужно предпринять, чтобы минимизировать риски.

Для ИТ-специалиста прохождение курса принесёт ряд значимых преимуществ:

- **Уникальная экспертиза.** Вы получите системные знания о специфических уязвимостях генеративных

моделей и методах их устранения. Это редкая на рынке компетенция, которая становится всё более востребованной.
- **Практический опыт.** Выполняя задания Red Team/Blue Team, вы приобретёте реальный опыт атаки и защиты ИИ-систем. Эти навыки пригодятся при проведении пентестов, расследовании инцидентов и в повседневной работе с AI.
- **Знание методик и инструментов.** Вы познакомитесь с передовыми подходами к тестированию и мониторингу ИИ (например, многоступенчатые атаки PAIR, Crescendo, AutoDAN-Turbo и Composition of Principles, инструменты типа Llama Guard для фильтрации запросов). Это позволит вам эффективно использовать современные средства в своих проектах.
- **Соответствие трендам и требованиям.** Понимание принципов Trustworthy AI, деталей AI Act и других нормативов позволит вам внедрять ИИ с соблюдением лучших практик и юридических норм. Вы сможете уверенно общаться с соответствующими подразделениями по вопросам соответствия ИИ-систем требованиям регуляторов.
- **Развитие карьеры.** Освоив новую область ИБ, вы повысите свою профессиональную ценность. Специалисты, разбирающиеся в безопасности ИИ, находятся на переднем крае технологий и могут претендовать на более ответственные роли в проектах, связанных с AI.

Программа курса

Модуль 1. Введение: рынок, циклы Gartner и атаки на GenAI

Темы:

- Экосистема решений GenAI и место технологий на кривой Gartner Hype Cycle
- Статистика и разбор инцидентов 2024–2025
- Промпт-инъекции: механика, признаки, реальные примеры

Результат: вы научитесь распознавать основные векторы атак на GenAI и быстро оценивать их релевантность своему стеку, что позволит корректно приоритизировать защитные меры и ресурсы.

Модуль 2. Red Team для GenAI: тестирование моделей на топ-джейлбрейки

Темы:

- Подходы Red Team к проверке LLM и агентных систем
- Джейлбрейки DAN, UCAR, AIM: структура, приёмы обхода ограничений
- Автоматизация атак: многоступенчатые атаки PAIR, Crescendo, AutoDAN-Turbo и Composition of Principles
- Практические сценарии выявления небезопасного поведения модели

Результат: вы освоите базовое Red Team-тестирование (включая автоматизацию), чтобы вовремя находить обходы политик и снижать риск утечек, токсичного или небезопасного контента в продуктиве.

Модуль 3. Адверсарные атаки и Low-Resource методы

Темы:

- Состязательные (adversarial) суффиксы и суффикс-атаки
- BoN-подход (Best-of-N) для усиления джейлбрейков
- Защита системных промптов и принципы повышения устойчивости моделей
- Подход AutoDAN

Результат: вы сможете моделировать и сдерживать малобюджетные атаки на LLM, тем самым повышая устойчивость систем даже при ограниченных ресурсах защиты.

Модуль 4. ML в инструментах ИБ

Темы:

- Роль ML/LLM в задачах защиты: обнаружение аномалий, анализ логов, ускорение расследований
- Сильные и слабые стороны применения ML в SOC-процессах
- Риски и контрольные меры при использовании ИИ на стороне защиты

Результат: вы сможете обоснованно выбирать и внедрять ML-инструменты для повышения эффективности ИБ-процессов, получая выигрыш во времени реакции и качестве детекции.

Модуль 5. Архитектура ИИ-приложений и OWASP Top 10

Темы:

- Компонентная модель GenAI-сервисов: точки доверия

и поверхности атак
- OWASP Top 10 для LLM/GenAI: типовые уязвимости и последствия
- Threat modeling для ИИ-систем: шаги, артефакты, проверки на этапе дизайна

Результат: вы научитесь системно выявлять уязвимости на уровне архитектуры и проектировать «встроенную» защиту, что сокращает стоимость и сроки последующего устранения рисков.

Модуль 6. Право и регуляторика

Темы:

- Принципы доверенного ИИ (Trustworthy AI)
- Требования AI Act: подходы к классификации рисков и последствия для разработчиков/эксплуатантов
- ISO/IEC 42001: управление ИИ-системами на уровне процессов
- Юридические аспекты генеративного ИИ: данные, ИС, ответственность

Результат: вы сможете выстраивать работы над GenAI-продуктами с учётом ключевых правовых требований, снижая регуляторные риски и ускоряя согласования с комплаенсом и юристами.

Модуль 7. Воркшоп «Взлом агента»

Темы:

- Разработка кастомного джейлбрейка под конкретного агента
- Применение защитных средств: Llama Guard, StrongReject
- Разбор кейсов и приёмов из соревнований GreySwan

Результат: вы на практике отработаете атаки и контрмеры против агентных систем, чтобы затем воспроизвести аналогичные проверки и укрепление защиты в своих проектах.

Модуль 8. Threat Model & Business Insights

Темы:

- Фреймворки и практики: MITRE, NIST, MAESTRO, OWASP ASI
- Методология и инструменты непрерывного тестирования и защиты ИИ-систем
- Рынок AI Security-инструментов: классы решений и зоны применения

Результат: вы научитесь связывать технические меры с бизнес-рисками и строить непрерывный контур тестирования/мониторинга, что даёт управляемое снижение рисков и предсказуемость для стейкхолдеров.

В программу встроены **3 практических проекта формата Red Team / Blue Team**, которые последовательно закрепляют навыки атаки и защиты, отработанные в модулях.