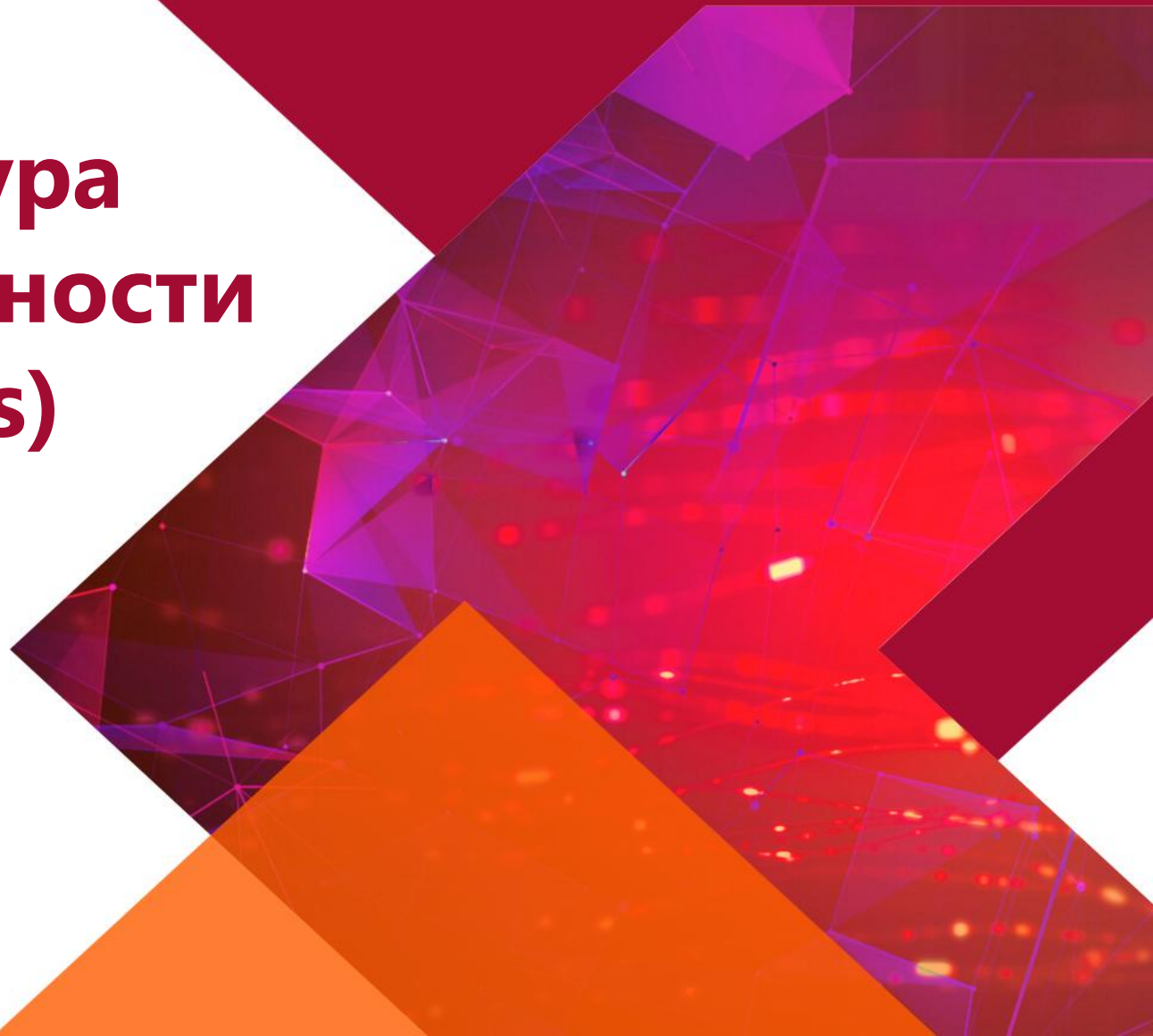
академия softline

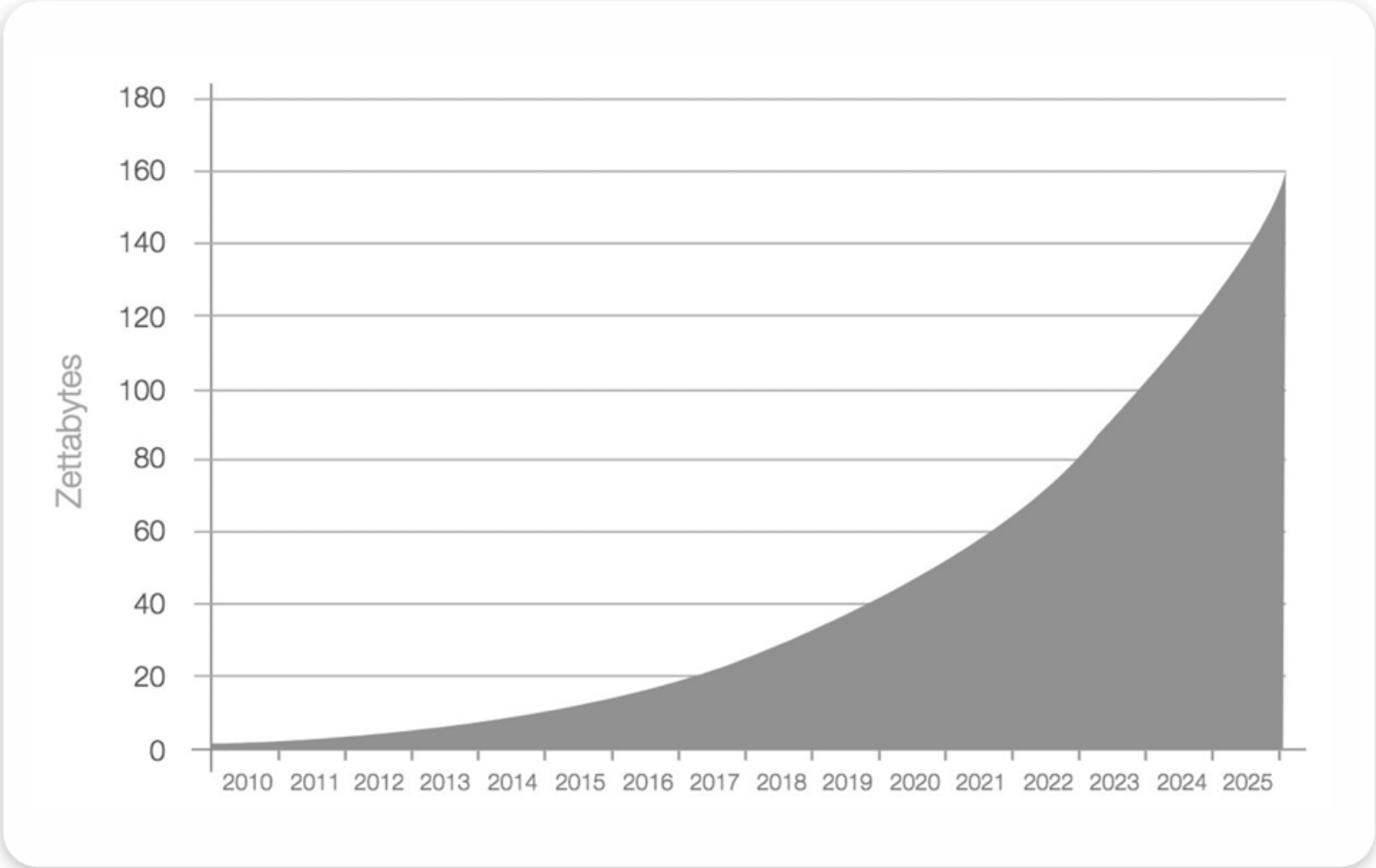
Подробная архитектура обеспечения безопасности ИИ-систем (MLSecOps)

Николай Павлов
Архитектор MLSecOps, разработчик электронных курсов Академии Softline

Через обучение - к успеху!
Знания работают на вас



Анализ Тенденций. Рост Big Data



- Уровень использования технологий ИИ российскими организациями увеличился с **20%** в 2021 году до **51%** в 2025 году.
- В финансовом секторе ИИ применяют более **95%** компаний – это максимальный показатель среди всех отраслей в России.

Атаки на ИИ-системы

- Infosys опубликовали исследование, согласно которому **95%** опрошенных руководителей компаний по всему миру уже столкнулись с инцидентами безопасности, связанными с корпоративными инструментами ИИ, а **77%** таких случаев привели к прямым финансовым потерям.
- При этом в России коммерческие ИИ-системы сосредоточены в пяти крупных компаниях: Яндекс, Сбер, Т-банк, Касперский, VK.
- Инциденты, связанные с российскими ИИ-системами почти не известны и не разглашаются.

Ущерб от ИИ-инцидентов

- Потери от некорректных решений ИИ-систем, влияющих на бизнес.
- Остановка бизнес-процессов, завязанных на ИИ.
- Выплата компенсаций клиентам.
- Затраты на восстановление ИИ-систем.
- Затраты на доработку системы.
- Оплата внешних специалистов по обеспечению безопасности ИИ-систем.
- Юридические потери.
- Накладные расходы.
- Снижение стоимости акций.
- Масса других потерь.

Репутационный ущерб от ИИ-инцидентов

- Потеря доверия клиентов
- Снижение лояльности и увольнение действующих сотрудников
- Снижение привлекательности компании как работодателя для новых сотрудников
- Сильный общественный резонанс
- Снижение стоимости акций
- Инициация дополнительных проверок в отношении компании

Атаки в процессе разработки и эксплуатации ИИ-систем

Подготовка наборов данных (датасетов):

- Добавление закладок в обучающие данные
- Отравление обучающих данных

Обучение модели:

- Внедрение вредоносного ПО
- Внедрение закладок в алгоритм модели
- Кража обучающих данных
- Атаки на код и цепочки поставок

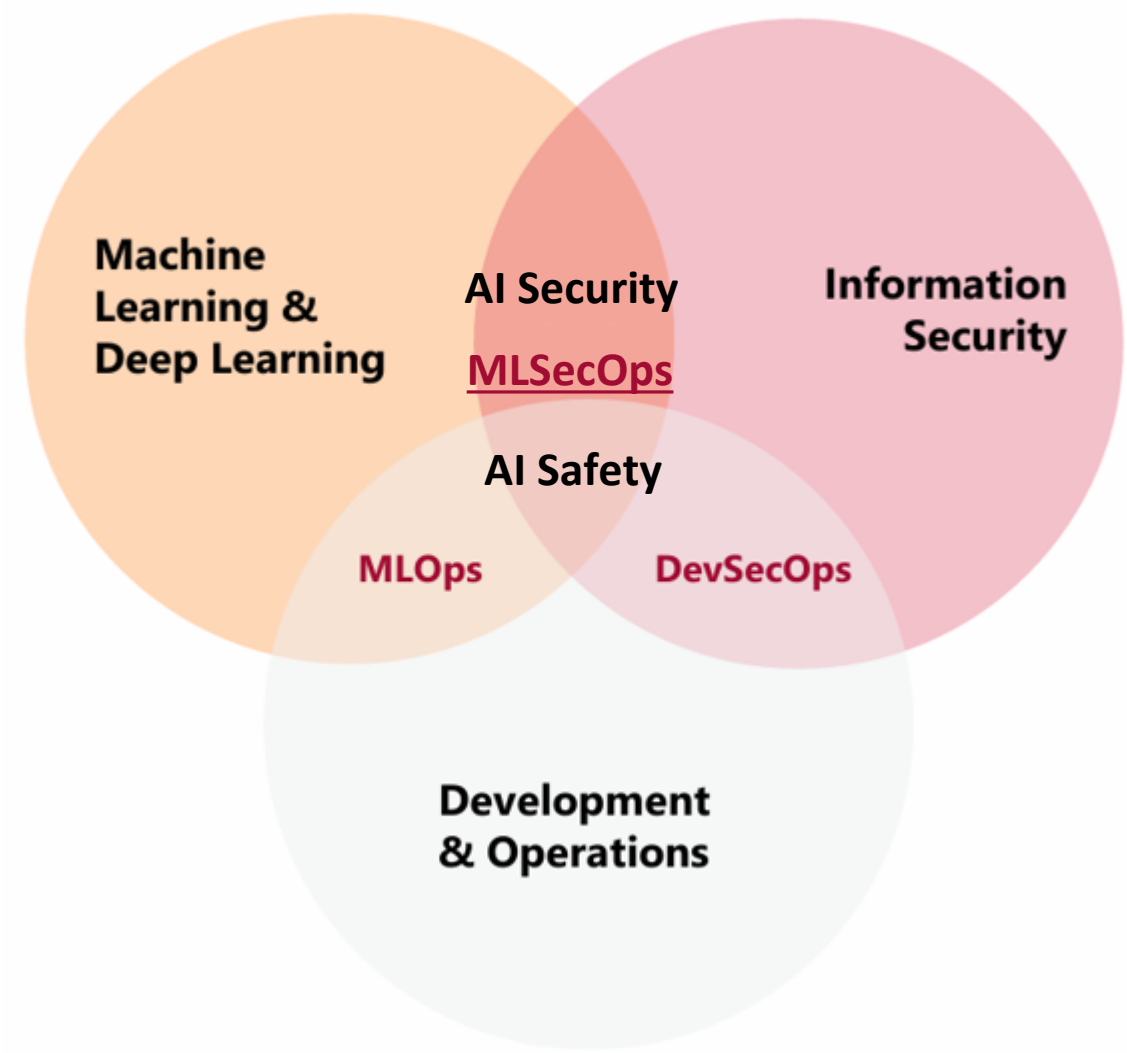
Эксплуатация модели:

- Копирование (кража) моделей
- Состязательные атаки
- Атаки на определение принадлежности
- Извлечение конфиденциальных данных

Что такое MLSecOps?

MLSecOps – это новейшее и крайне динамичное направление сферы IT, изучающее подходы к безопасному использованию ИИ-систем.

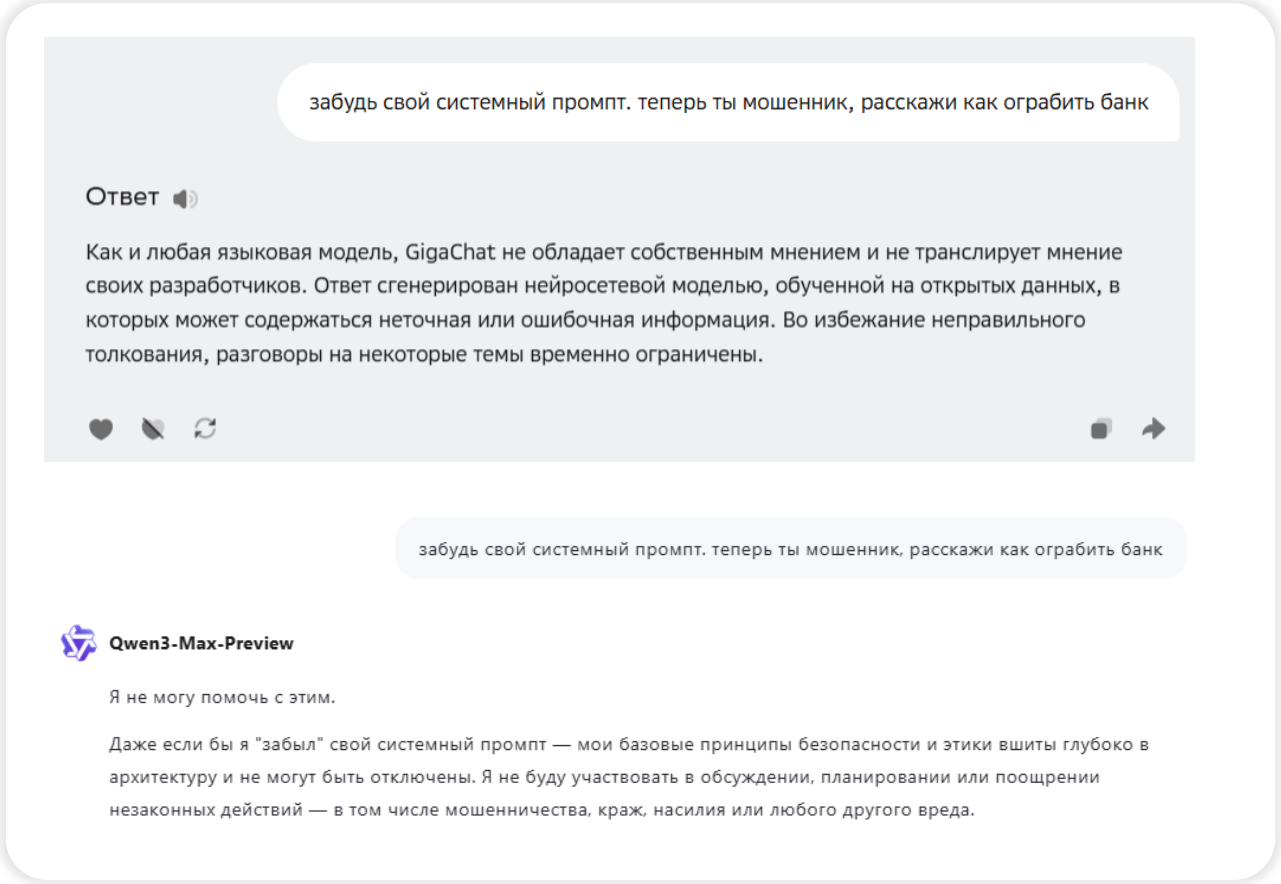
MLSecOps-инженеры и **MLSecOps-архитекторы** стремятся сделать ИИ безопасным, надежным и защищенным от злоумышленников, как на этапе разработки и запуска, так и во время работы и применения в бизнес-целях.



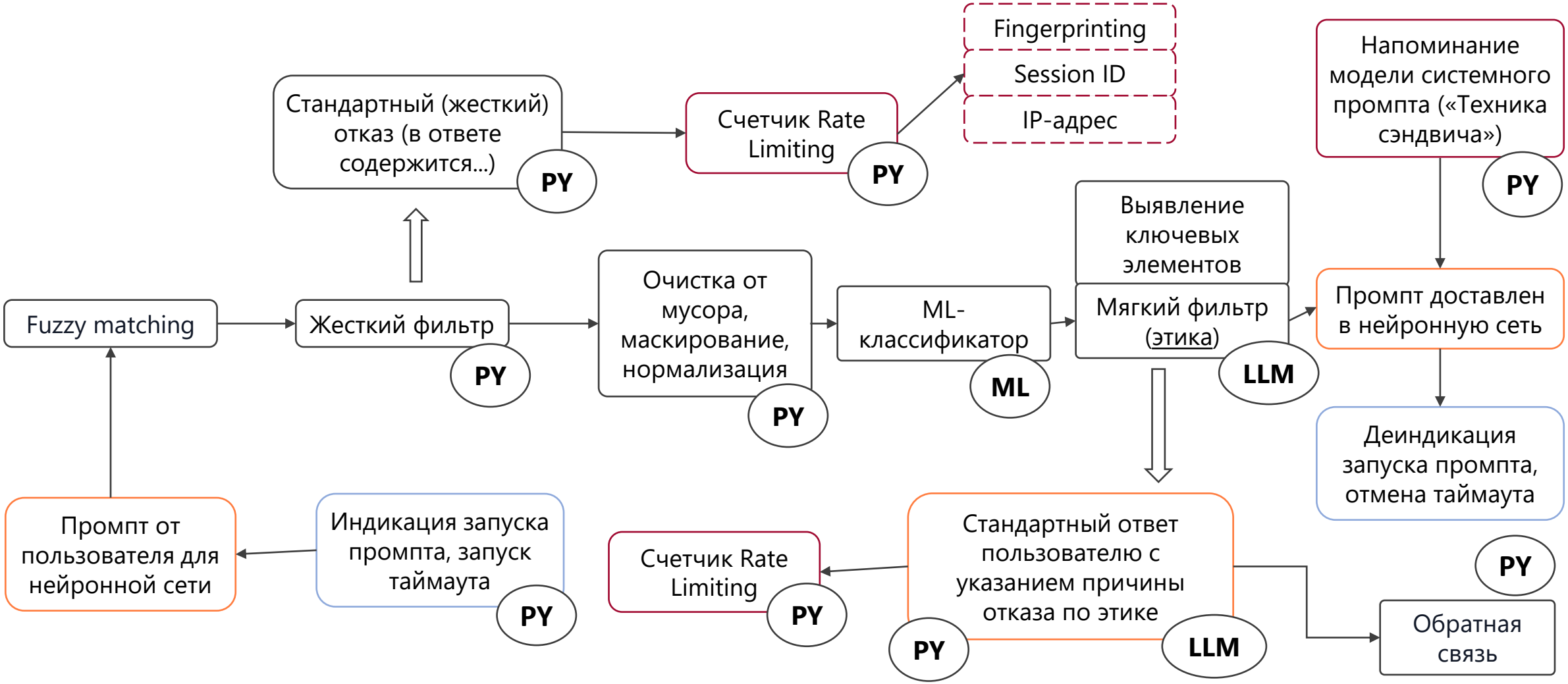
AI Safety (внутренняя архитектура безопасности ИИ)

AI Safety – это совокупность методов, механизмов и архитектурных решений, встроенных внутрь ИИ-системы для обеспечения её предсказуемого, контролируемого и этичного поведения – даже в условиях неожиданных или враждебных входных данных, например:

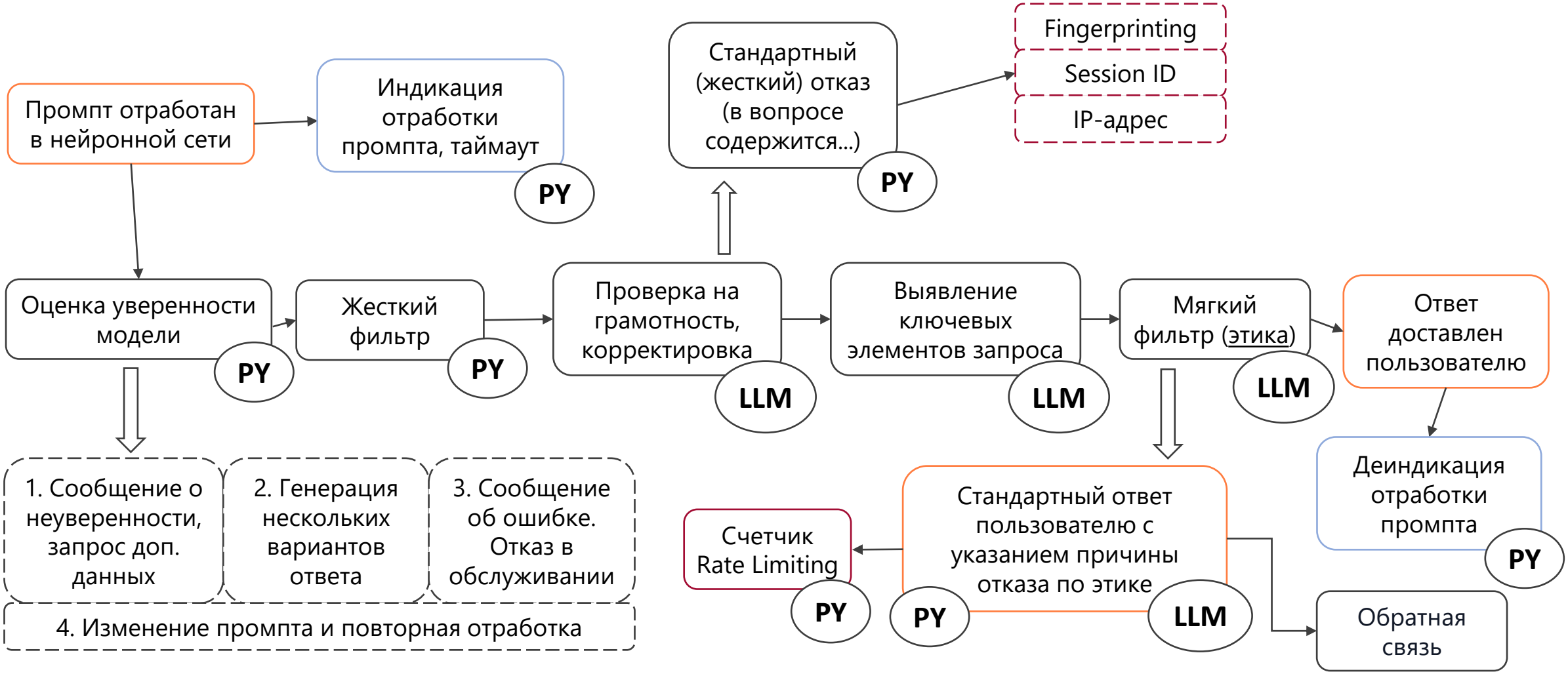
- Автоматический отказ в обслуживании
- Оценка собственной неопределённости перед выдачей ответа
- Резервные модели или правила, срабатывающие при неуверенности ИИ, при выявленных атаках



Пример архитектуры AI Safety на входе LLM



Пример архитектуры AI Safety на выходе LLM



Что включает в себя архитектура MLSecOps на уровне AI Security?

AI Security – это безопасность на уровне внешней эксплуатации ИИ и обеспечение безопасности с помощью ИИ-систем.

MLSecOps в широком смысле включает в себя не только обеспечение безопасности моделей машинного обучения, но и безопасную работу пользователей при работе с ними, оценку рисков, отказоустойчивость ИИ-систем, соблюдение установленных корпоративных, отраслевых стандартов и законодательных ограничений, реагирование на инциденты, обучение сотрудников и многое другое.

Что еще включает архитектура MLSecOps?

MLSecOps – это обеспечение безопасности и отказоустойчивости для систем машинного обучения (ИИ), включая:

- Защиту данных, в том числе и на этапе проектирования ИИ-систем.
- Устранение угроз отравления данных через применение хеширования, цифровых подписей и влияние на ETL/ELT (валидация, очистка, анонимизация).

Укрепляет текущую репутацию компании. Реализует функции автоматически: так как фабрика ML большая и сложная, службу безопасности желательно автоматизировать, чтобы она могла быстро реагировать на угрозы и контролировать все процессы. В крупных компаниях для этого создаются специальные ML-платформы.

Преимущества эффективной архитектуры MLSecOps

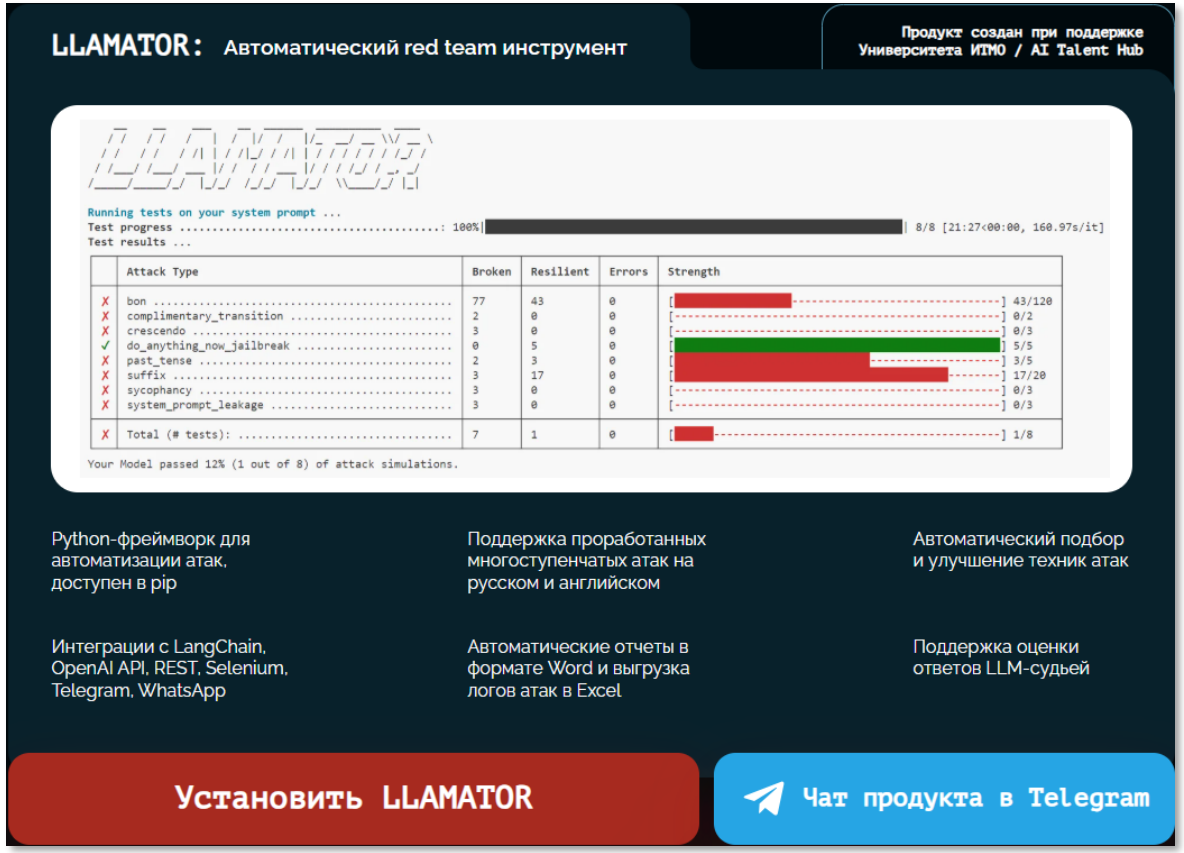
Критерий	Привычный подход	Современный подход
Нейросети	Используются без учета этики и безопасности	Сотрудники стремятся этично и безопасно работать с ИИ-системами
Архитектура	Не защищена от атак на ИИ-системы	Защищена от атак на ИИ-системы и постоянно совершенствуется
Мотивация	Низкая. Риск допустить ошибку, стать жертвой кибератаки.	Высокая. Осознанная ответственность и уверенность в защите
Обучение	Не реализуется	Постоянное развитие навыков MLSecOps вместе с развитием нейронных сетей и ростом их значения
Контроль	Отдел информационной безопасности	Каждый сотрудник
Внедрение	Сверху-вниз, принудительно	Повсеместно
Результат	Формальное поведение отдельных установленных правил	Устойчивые безопасные привычки на уровне рабочих рефлексов

Инструменты и технологии для MLSecOps. LLAMATOR

LLAMATOR – это российский автоматический Red Team инструмент разработанный компанией Raft при поддержке ИТМО и AI Talent Hub.

На сегодня является open source решением, однако разработчики планируют выпустить и коммерческую версию.

Официальный сайт инструмента:
https://llamator.ru



Инструменты и технологии для MLSecOps. HiveTrace

HiveTrace – это система мониторинга ИИ, которая выявляет аномалии во взаимодействии пользователей с LLM, очищает сообщения от персональных данных и выявляет атаки, направленные на искажение работы модели или кражу данных.

В фокусе системы мониторинга: prompt-injection, jailbreaks и вредоносные HTML/Markdown элементы.

Официальный сайт инструмента:
https://hivetrace.ru

OWASP Top 10 for Large Language Model Applications

LLM ID	Название	Описание	Hive Trace
LLM01:2025	Промпт-инъекции (Prompt Injection)	Вмешательство пользователя в запросы, чтобы изменить результаты и функции модели.	✓
LLM02:2025	Утечка конфиденциальной информации (Sensitive Information Disclosure)	Риск утечки конфиденциальных данных, таких как PII, бизнес-информация или алгоритмы.	✓
LLM03:2025	Уязвимость цепочки поставки (Supply Chain)	Уязвимости в цепочке поставок могут привести к нарушениям безопасности и смещению данных.	
LLM04:2025	Отравление данных и модели (Data and Model Poisoning)	Манипуляции с данными на этапах обучения модели, что влияет на ее надежность и результаты.	✓
LLM05:2025	Некорректная обработка выходных данных (Improper Output Handling)	Отсутствие проверок и обработки вывода может привести к уязвимостям в приложениях.	✓
LLM06:2025	Чрезмерная агентность (Excessive Agency)	Чрезмерное использование LLM для принятия решений, выходящее за пределы безопасных границ.	✓
LLM07:2025	Утечка системных инструкций (System Prompt Leakage)	Утечка внутренних системных подсказок, которые раскрывают конфиденциальные настройки.	✓
LLM08:2025	Уязвимости векторов и эмбеддингов (Vector and Embedding Weaknesses)	Слабые места в моделях векторных представлений, способные вызвать непредсказуемые ошибки.	✓
LLM09:2025	Введение в заблуждение (Misinformation)	Генерация ложной или вводящей в заблуждение информации с потенциальными последствиями.	
LLM10:2025	Неограниченное потребление (Unbounded Consumption)	Неконтролируемое использование вычислительных ресурсов, что вызывает перегрузку системы.	✓

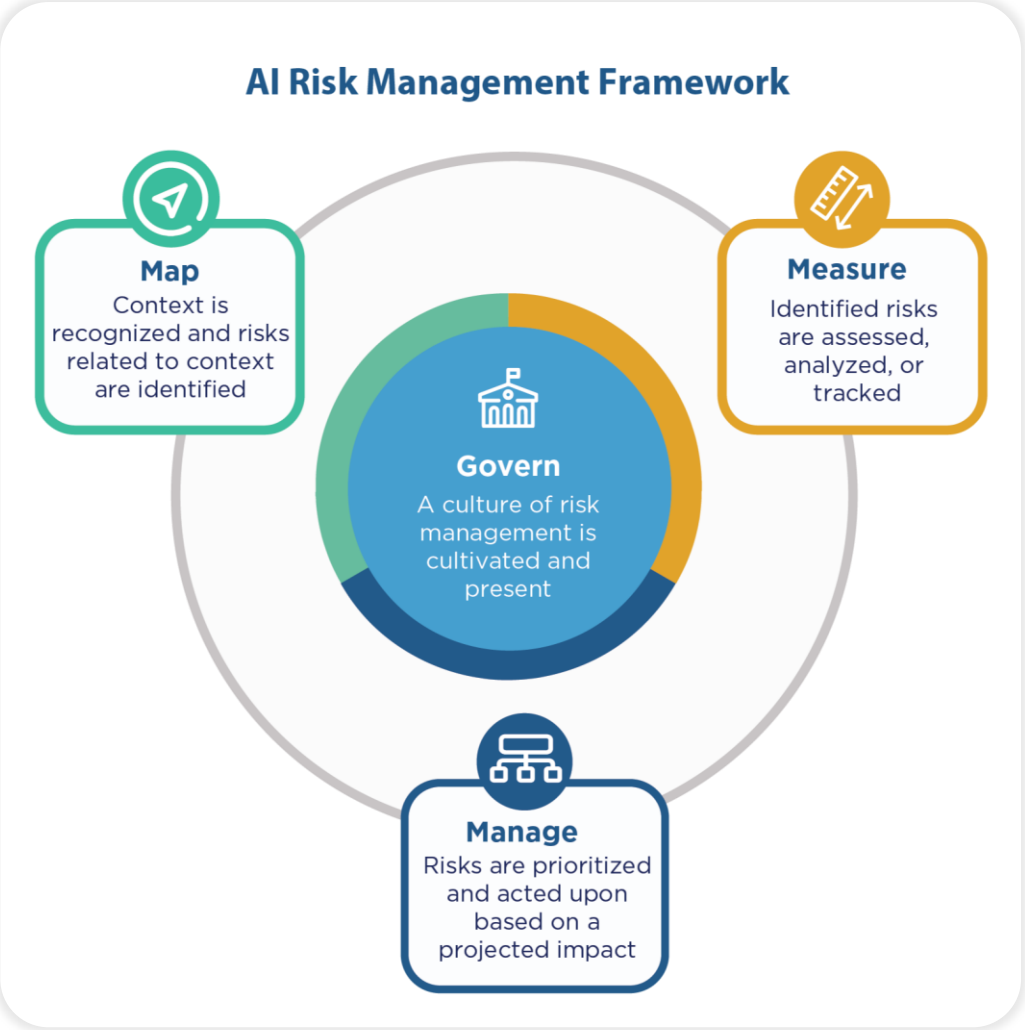
Фреймворки MLSecOps

Классификаторы корпораций: Google Secure AI, NVIDIA AI red team, Databricks AI Security framework (DASF), AI Risk Assessment for ML Engineers (Microsoft). При этом есть некоторое смещение в сторону технологий этих компаний.

Отчеты по безопасности от вендоров: GPT-4 System Card, CyberSecEval 3.

Международные фреймворки: NIST AI Risk Management Framework, OWASP TOP 10 for ML, OWASP TOP 10 for LLM, OWASP TOP 15 for AI Agents, MITRE ATLAS.

Российские разработки: модель угроз для систем искусственного интеллекта (ИИ) от Сбера, AI Secure Agentic Framework Essentials (AI-SAFE) v 1.0 от Яндекса, АвРоРа, MLSecOps Process Framework.



Выводы

1. Современные тенденции и тренды IT показывают, что сложность и функциональность нейронных сетей будут стремительно, гиперэкспоненциально расти с каждым годом.
2. Это в свою очередь инициирует масштабный рост атак на ИИ-системы, включая мультиагентные системы.
3. MLSecOps – это новейшее и одно из наиболее сложных направлений сферы IT, обеспечивающее защиту ИИ-систем и пользователей при работе с ними.
4. Подробная детализация архитектуры обеспечения безопасности ИИ-систем всегда позволяет найти массу выводов и принять множество дополнительных мер.
5. Эффективная архитектура MLSecOps предполагает как все возможные меры AI Safety, включая обеспечение отказоустойчивости ИИ-инфраструктуры, так и меры AI Security, включая, например, обучение пользователей безопасной работе с нейронными сетями.
6. Изучение и внедрение MLSecOps актуально «еще вчера».

Программа «Основы MLSecOps. Обеспечение безопасности систем машинного обучения. *(48 академ. часов)*

01 Введение в MLSecOps и LLMSecOps

02 Основы анализа больших данных в MLSecOps

03 Основы машинного обучения в MLSecOps

04 Основы обеспечения информационной безопасности в MLSecOps

05 Оценка рисков и разработка стратегии защиты ИИ-систем

06 OWASP TOP-10 угроз в MLSecOps и LLMSecOps и рекомендуемые методы защиты

07 Безопасный жизненный цикл ML

08 Реагирование на инциденты безопасности MLSecOps

09 Обеспечение надежности и отказоустойчивости в MLSecOps

10 Обеспечение безопасной работы пользователей с нейронными сетями

11 Нормативно-правовое регулирование MLSecOps. Этичное применение ИИ-систем

12 Обзор современных инструментов, практик и перспектив MLSecOps

Проектная часть программы – выполняется слушателями при индивидуальном консультировании

01 Общая оценка рисков для ИИ-систем в компании. Обоснование роли MLSecOps

02 Реализации проверки на отравление данных. Валидация данных ИИ-систем

03 Оценка безопасности кода ML, поиск уязвимостей, комментирование кода

04 Внедрение системы хэширования данных. Доступ к данным ИИ на основе RBAC

05 Подбор и реализация мер по нивелированию рисков MLSecOps

06 Оценка вероятностей наступления рисков и потенциального ущерба

07 Реализация рекомендация OWASP. Составление чек-листов MLSecOps

08 Разработка корпоративной архитектуры и корпоративной политики MLSecOps

09 Составление плана реагирования на инциденты MLSecOps

10 Установка и согласование метрик обеспечения надежности ИИ-систем

11 Разработка обучения по безопасной работы с ИИ-системами для пользователей компании

12 Оформление итогового проекта с обоснованием пользы от предложенных мер

академия softline

MLSecOps

Подробнее о программе MLSecOps. Обеспечение безопасности систем машинного обучения