

// вебинар CyberED

Безопасная разработка в эпоху ИИ-агентов

Как ИИ-агенты меняют поверхность атаки — и как
строить безопасный жизненный цикл разработки (ADLC).

20 августа • 19:00 МСК

// досмотри до конца — будет бонус на обучение



// спикеры

Кто ведёт



Денис Макрушин

Директор по продуктам безопасной разработки

Яндекс

- развивает платформу и сервисы безопасной разработки
- исследует применение ИИ для поиска и исправления уязвимостей
- ранее был техническим директором по кибербезопасности в МТС
- руководил технологическим центром Huawei; ранее исследовал угрозы и уязвимости в Kaspersky GReAT



Глеб Михеев

Chief Product Officer ГигаАгента

SBER BANK

- развивает генеративный ИИ и GigaChat
- занимается агентной разработкой и мультиагентными системами
- работает в коммерческой разработке с 2003 года
- руководит программным комитетом AgenticDevConf

// как меняется работа разработчика

Усилия смещаются вверх по стеку

2024

ДИЗАЙН И АРХИТЕКТУРА
Системы, инфраструктура, решения

КОД РУКАМИ
Строка за строкой, функция за функцией

РЕВЬЮ И ДЕБАГ
PR, тесты, чиним сломанное

AI-GENERATED CODE



2027

ДИЗАЙН И АРХИТЕКТУРА
Системы, инфраструктура, как сервисы общаются

УПРАВЛЕНИЕ ИИ-АГЕНТАМИ
Промпты, контекст, ограничения

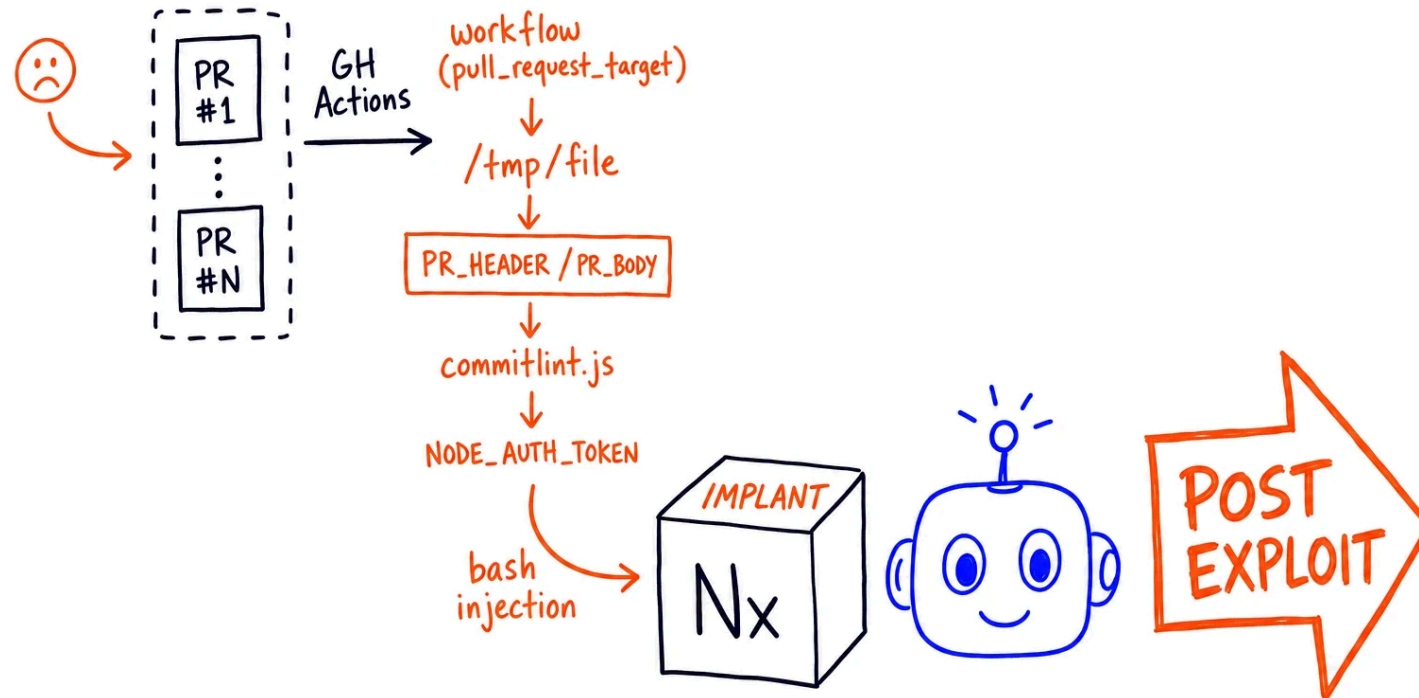
ВЕРИФИКАЦИЯ AI-КОДА
Ревью того, что написали машины

Как меняется безопасность разработки в эпоху ИИ-агентов?

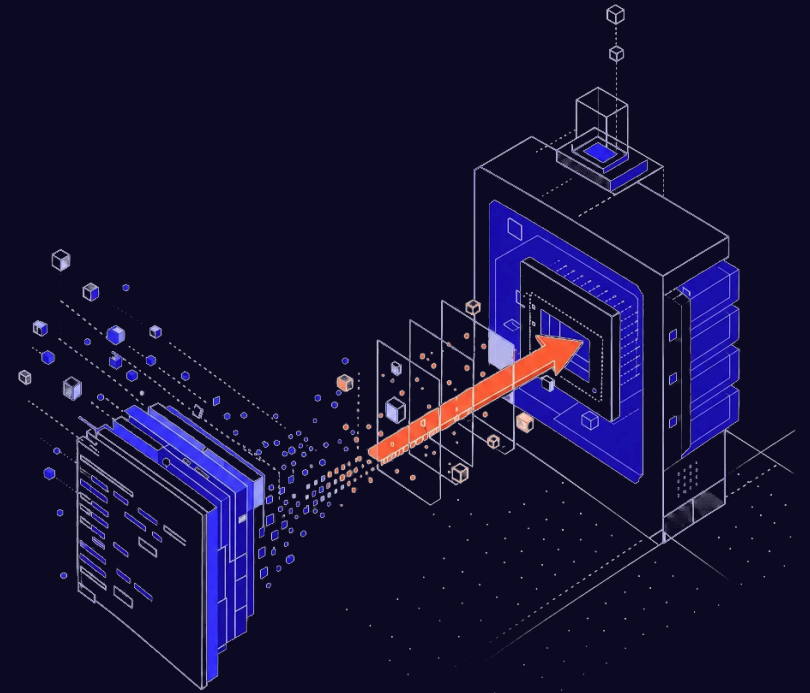


// реальный кейс · атака на цепочку поставки

Недоверенный PR → выполнение в CI → кража токена → имплант

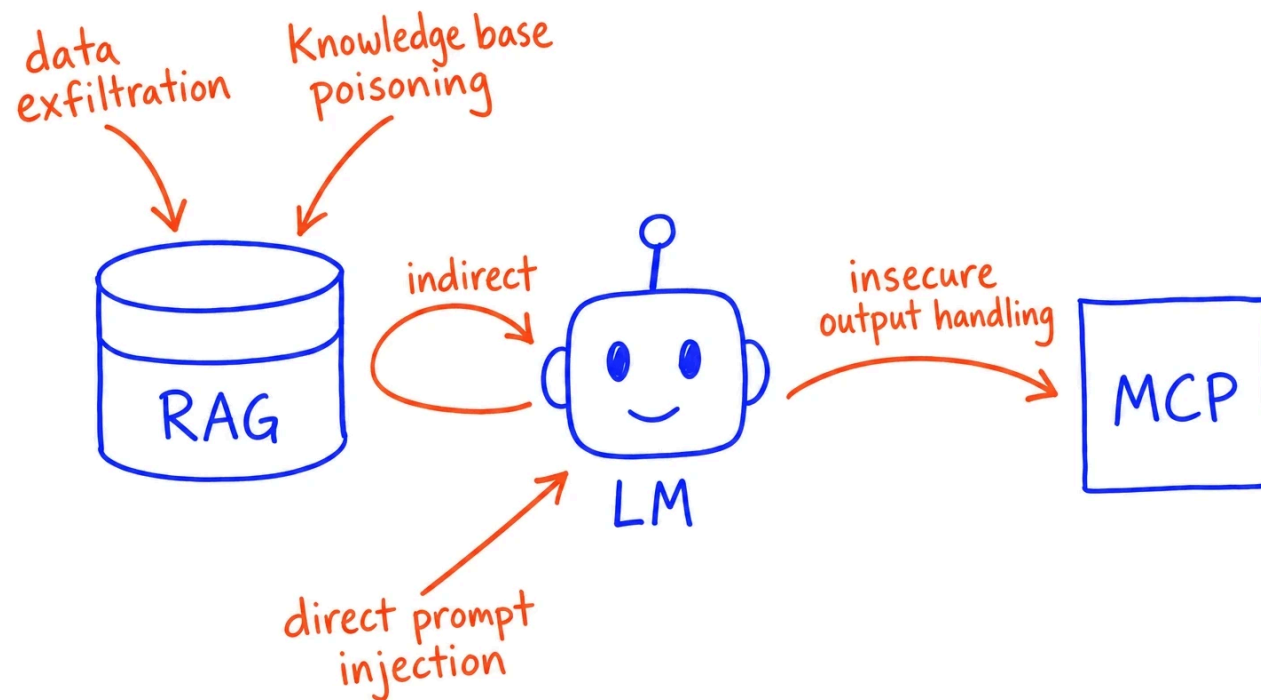


Как недоверенные данные
становятся **инструкциями**?

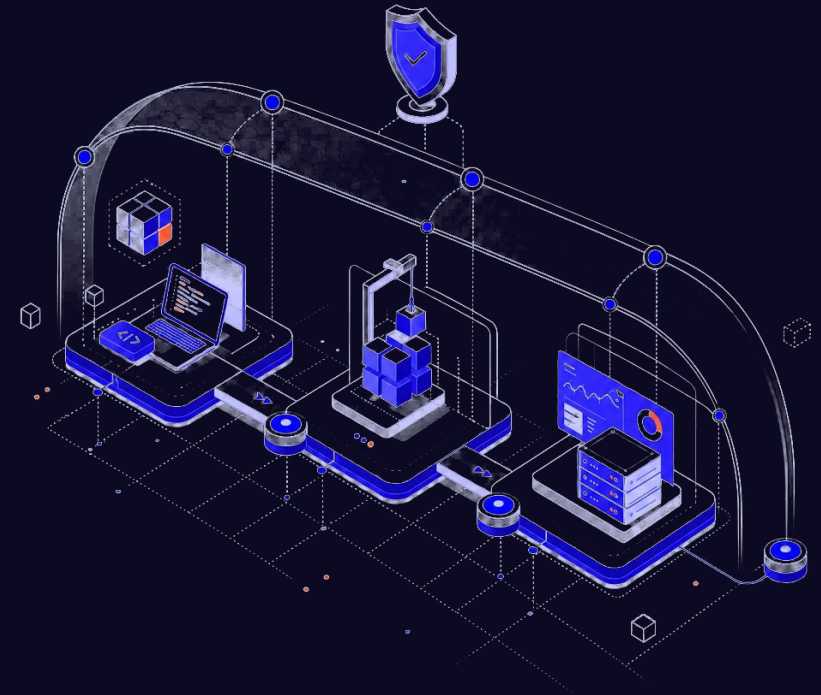


// поверхность атаки ИИ-агента

Компоненты и векторы атак



Что делать? Как строить безопасный ADLC?



// как проверять ИИ-системы на прочность

AI Red Teaming

Систематическая проверка LLM и агентов: prompt injection, утечки, обход политик, небезопасные действия инструментов.



// AI Red Teaming · tooling

Инструменты AI Red Teaming

✕ **LM Eval Harness**

оценка моделей

✕ **CleverHans**

adversarial

✕ **ART**

adversarial
robustness

✕ **Giskard**

тестирование LLM

✕ **CyberSecEval**

безопасность кода

✕ **PyRIT**

red teaming

✕ **EasyEdit**

правка знаний

✕ **Promptmap**

prompt injection

✕ **FuzzyAI**

фаззинг LLM

✕ **Fickling**

pickle /
сериализация

✕ **Rigging**

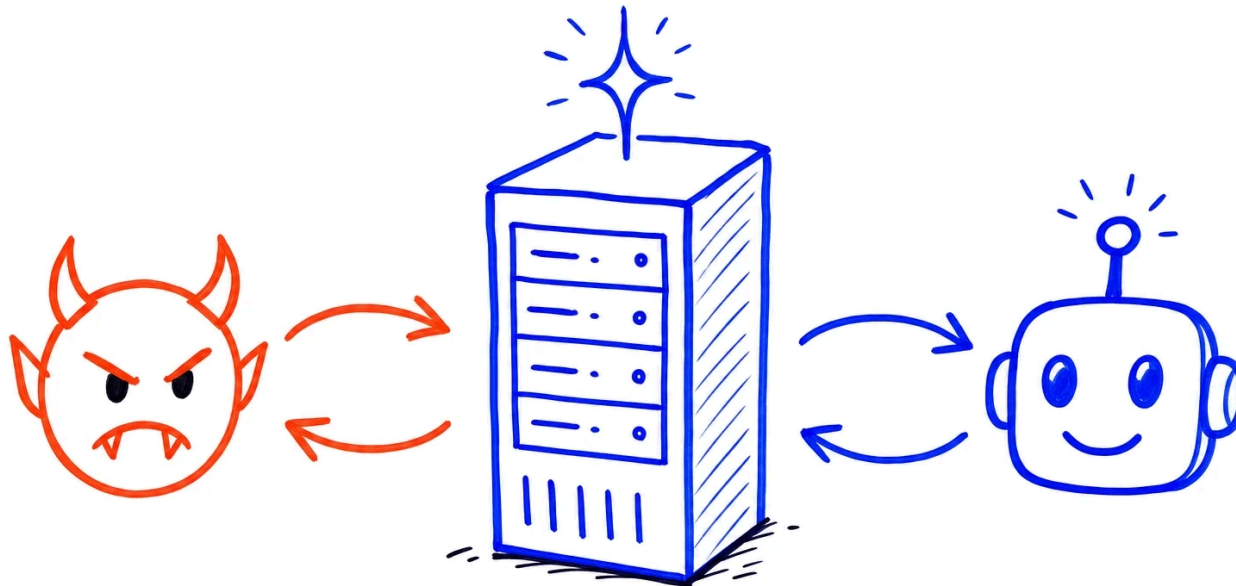
агентный фреймворк

✕ **Judges**

LLM-as-judge

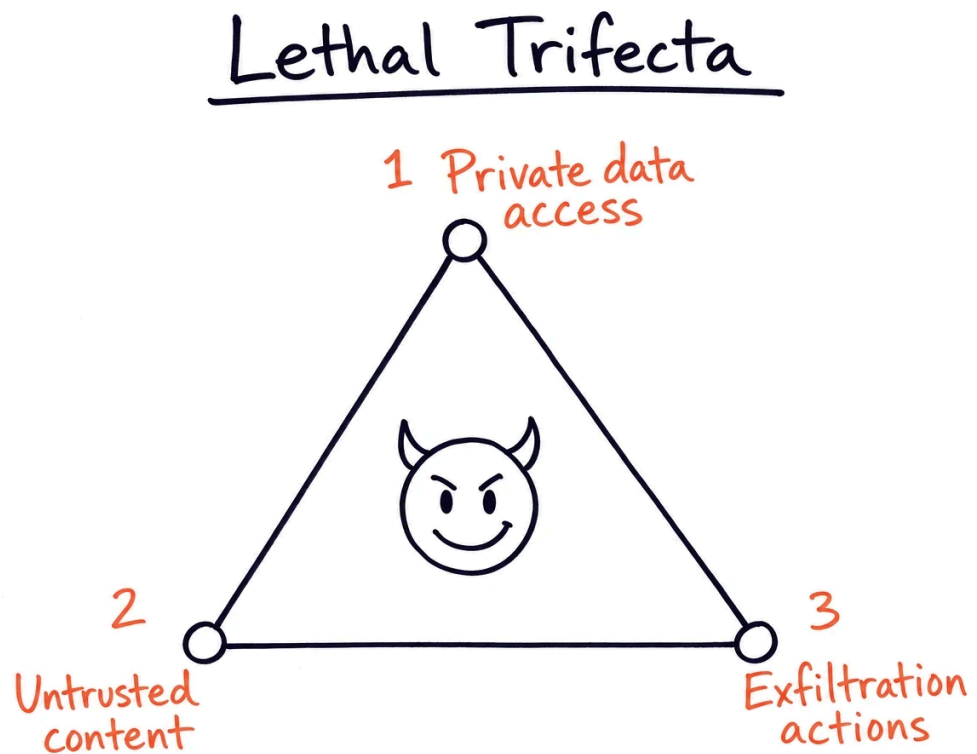
// агент выходит во внешний мир

Атакующий и агент — по разные стороны одной системы



// три условия ущерба — сходятся вместе

Lethal Trifecta



// shift down

Обеспечение качества сдвигается на платформу

сегодня · всё на разработчике

Обеспечение качества и безопасности

Разработка фич

develop · deploy · operate



shift
down

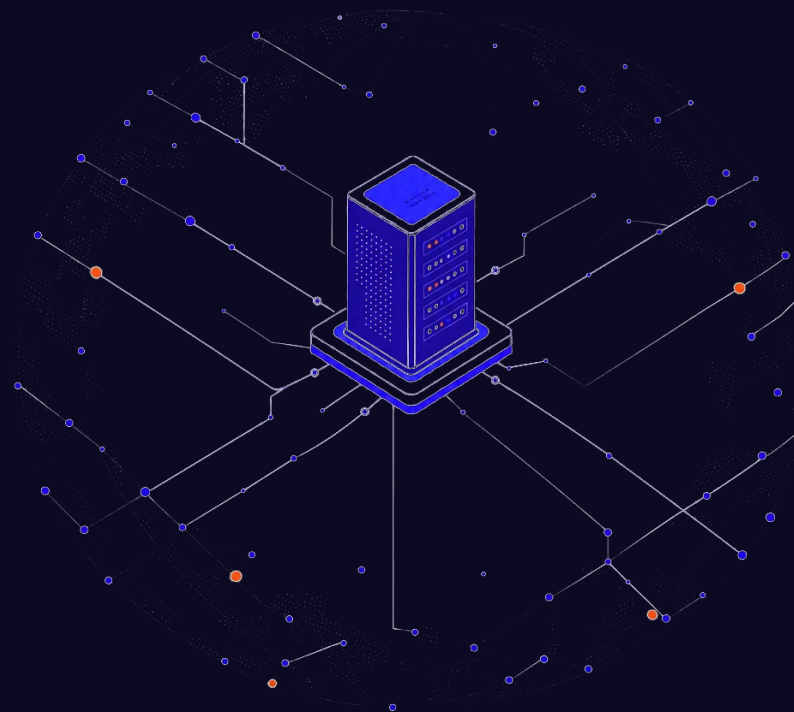
shift down · платформа берёт качество

Разработчик — только разработка фич

Платформа — обеспечение качества и безопасности

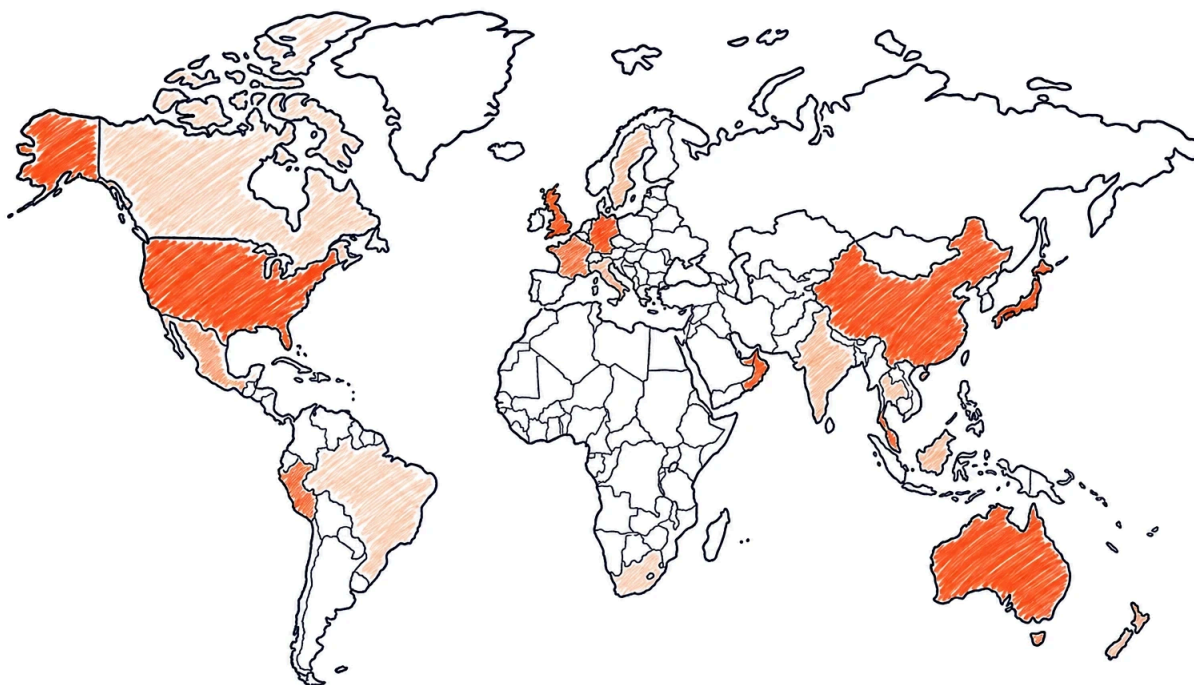
меньше рутины на инженере · контроль — в платформе

А если агент доступен
снаружи?
поговорим про МСР



// мы просканировали интернет

Открытые MСР-серверы — по всему миру



1677 обнаружено

585 подтверждено

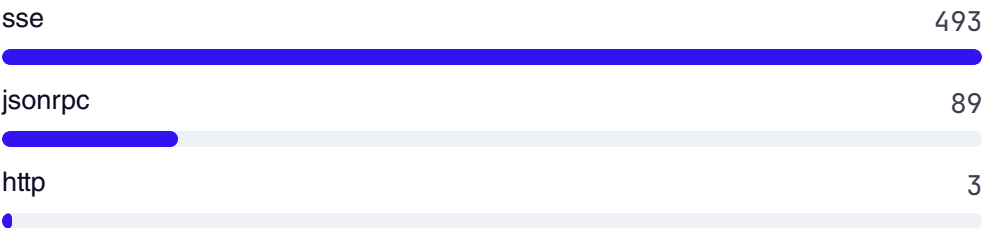
568 критических

// скан открытых МСР-серверов

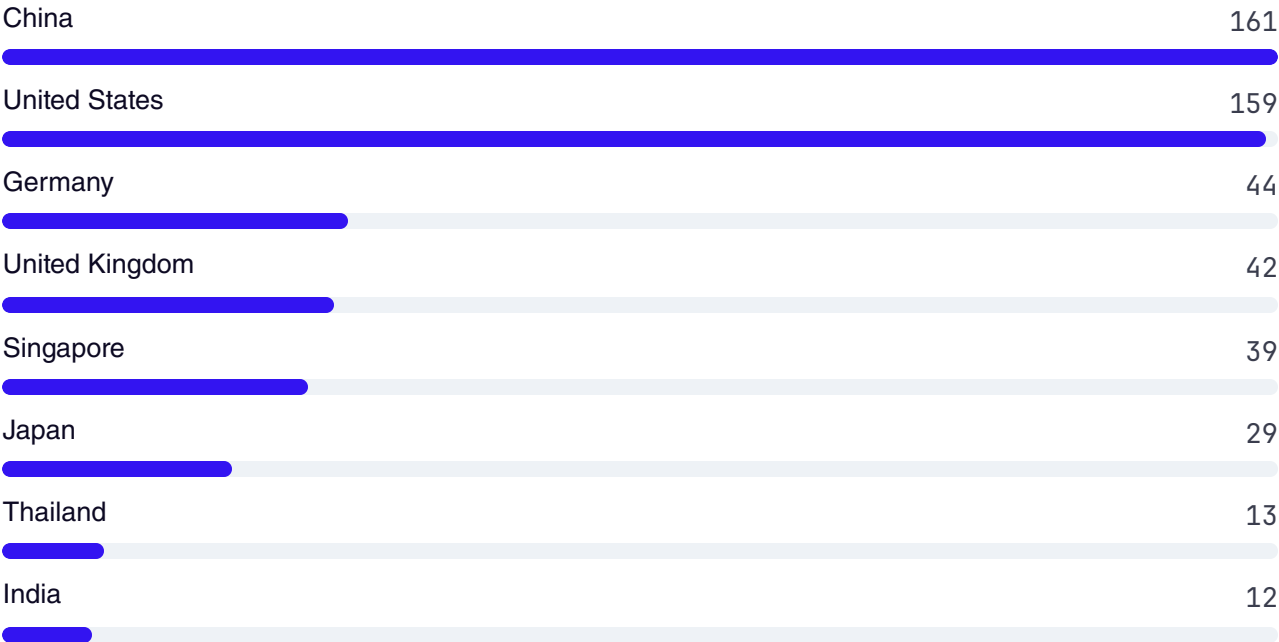
585 подтверждено · 568 критических



// transport type



// top countries



// как это выглядит в логах

Массовое сканирование MCP-эндпойнтов

```
{ ... "HTTPMethod":"POST","RequestURI":"/mcp","Description":"Apache 401","SourceIp":"202.120.234.94","SourcePort":"49180" }  
{ ... "HTTPMethod":"POST","RequestURI":"/mcp/","Description":"Apache 401","SourceIp":"202.120.234.94","SourcePort":"49180" }  
{ ... "HTTPMethod":"POST","RequestURI":"/api/mcp","Description":"Apache 401","SourceIp":"202.120.234.94" }  
{ ... "HTTPMethod":"POST","RequestURI":"/sse","Description":"Apache 401","SourceIp":"202.120.234.94" }  
{ ... "HTTPMethod":"POST","RequestURI":"/sse/","Description":"Apache 401","SourceIp":"202.120.234.94" }  
{ "clientInfo":{"name":"gitmc-org-mcp-scanner","version":"1.0.0"}, "HTTPMethod":"POST","RequestURI":"/mcp" }
```

сканер бьёт по /mcp, /api/mcp, /sse — ищет открытые серверы по всему интернету

// от уязвимости до эксплуатации

Time-to-Exploit схлопывается

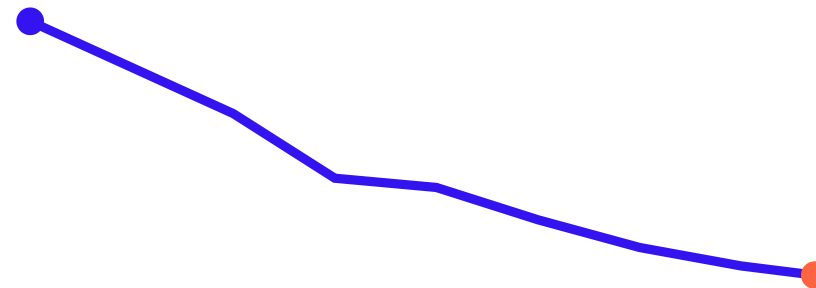
2018

2.3 года



2026

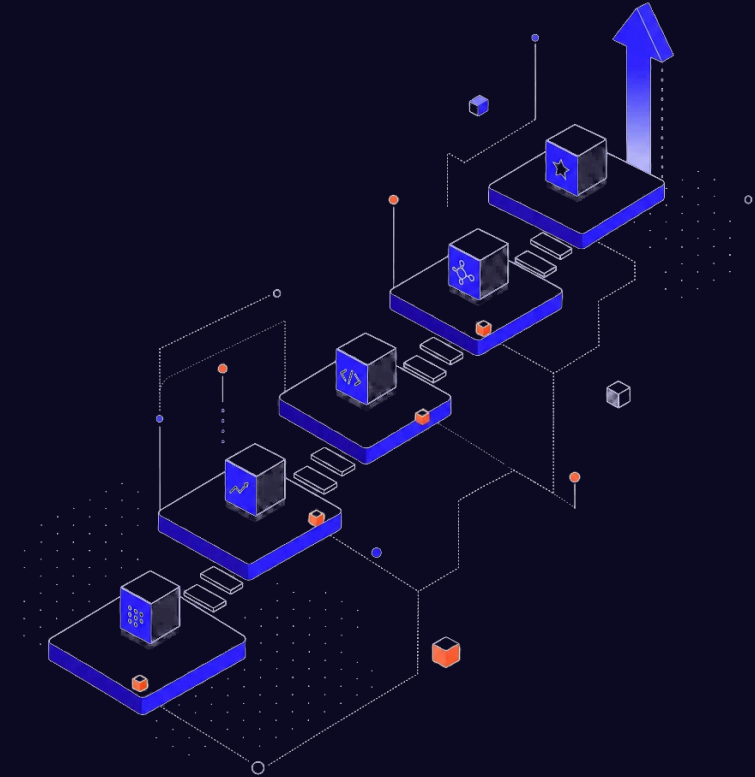
1.6 дня



Mean TTE · 3 515 CVE-exploit map (CISA KEV, VulnCheck)

Окно на реакцию сжалось с лет до дней — защита должна быть встроена в процесс, а не догонять.

Какие **навыки** становятся
актуальны в контексте **ADLC**?



// итог • и что дальше

Безопасность ИИ — это про **контроль**, не запрет

Границы доверия, ограничение полномочий, security-evaluations и мониторинг — на каждом шаге ADLC.

// и этому можно научиться

Интенсив «Безопасность ИИ-систем и агентов»



// интенсив CyberED · AI Security Engineering

Безопасность ИИ-систем и агентов

32

АК.Ч ВСЕГО

4

СЕМИНАРА С ПРЕПОДАВАТЕЛЕМ

11

АК.Ч С ПРЕПОДАВАТЕЛЕМ

21

АК.Ч САМОСТОЯТЕЛЬНО

// по итогам участник получит

- > Защищённый прототип ИИ-агента
- > Модель угроз и архитектурную схему системы
- > Матрицу разрешённых действий и ограничения
- > Security-evaluations, release-гейты и AI Security Framework

// документ

Удостоверение о повышении квалификации

официальный документ по итогам интенсива

// программа · 4 занятия · практика и результат

Путь атакующего и защитника

занятие 1 Архитектура, RAG и модель угроз	практика Проанализировать архитектуру и воспроизвести атаку через недоверенный источник RAG	результат Архитектурная схема, модель угроз, доказательство атаки и повторной проверки
занятие 2 Атака и защита ИИ-агентов и MCP	практика Воспроизвести indirect prompt injection; добавить validation, sandbox, approval, журналирование; повторить атаку	результат Защищённый MCP-инструмент, матрица разрешённых действий, результат повторного теста
занятие 3 Secure AI SDLC и Security Evaluations	практика Превратить сценарии угроз в security-evaluation, применить защиту, определить release-гейты	результат Eval-набор, результаты до/после усиления, перечень security gates
занятие 4 Runtime Security, IR и AI Governance	практика По журналам обнаружить и расследовать опасное поведение агента, восстановить систему	результат План мониторинга, реагирования, восстановления и повторного допуска

// кому подойдёт и зачем

Целевая аудитория

AppSec-инженеры · Secure SDLC

Адаптировать AppSec-практики к разработке и внедрению ИИ-систем

DevSecOps- и DevOps-инженеры

Безопасно встроить ИИ-инструменты в CI/CD и эксплуатацию

Архитекторы решений

Проектировать безопасную архитектуру, доступы и защиту ИИ-компонентов

Разработчики и Security Champions

Безопасно применять ИИ-ассистентов и учитывать риски при проектировании

CISO и Security Leads

Сформировать требования к AI Security и AI Governance

Инженеры ИБ · будущие AI Security

Получить базовые компетенции по защите ИИ-систем и MLSecOps

// спецпредложение

Забери интенсив со скидкой

С 1 сентября — полная цена. В августе действует спеццена, а если забронируешь **сегодня** — дополнительная скидка 5 000 ₽.

с 1 сентября

~~60 000 ₽~~

в августе · -25%

45 000 ₽

бронь сегодня · ещё -5 000 ₽

40 000 ₽

Получи все материалы программы у менеджера

отсканируй QR · цена 40 000 ₽ действует только сегодня

+ в подарок в августе: при покупке — второй интенсив «AI-усиленный пентест» от Сергея Зыбнева бесплатно