



Слепая зона SOC: AI-агенты и LLM

Как ускорить разбор инцидентов
и взять LLM-контур под контроль



Илья Мельников

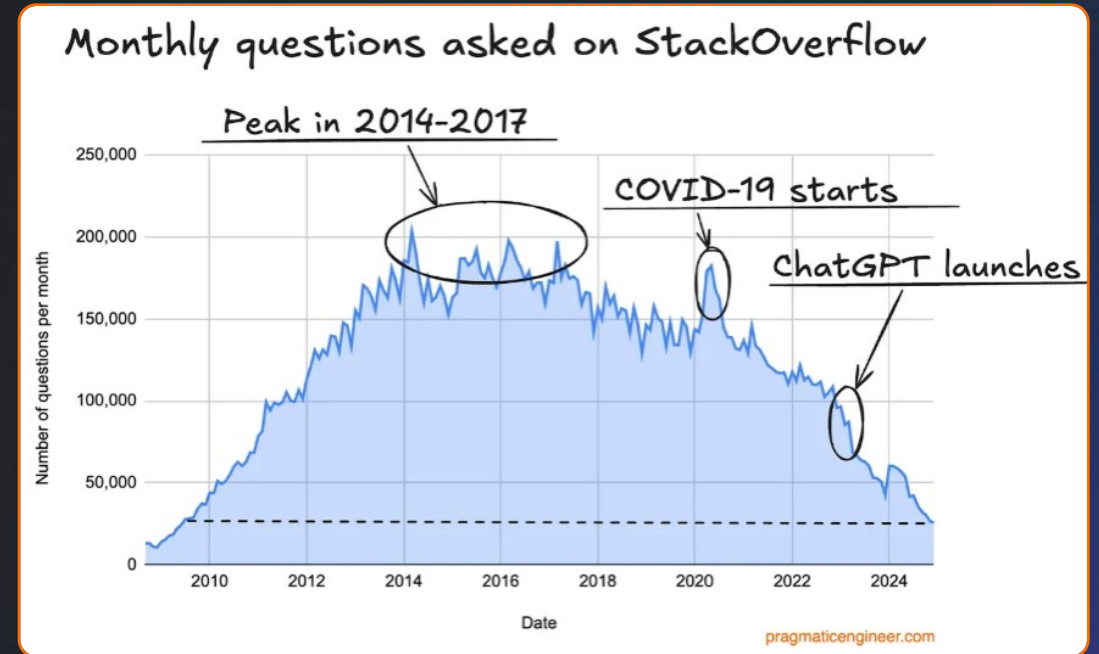
Эксперт по мониторингу, ВолгаБлоб



Нельзя игнорировать технологии, которые меняют правила игры



Популярность запросов по теме AI



Количество вопросов на StackOverflow

AI ускоряет работу — и сам становится целью атак

AI как инструмент



Разбор инцидента быстрее



Поиск по смыслу, а не гегехр



Рутина L1 в пресетах



Сводки без ручной сборки

AI как объект атаки



Инъекции промптов и утечки



Агенты с избыточными правами



Расходы и нагрузка вне контроля



Теневые внешние LLM

Усиливаем SOC и берём LLM-контур под контроль

ЗАДАЧА 1 · УСКОРИТЬ РАЗБОР ИНЦИДЕНТОВ



01

Работа с данными

Семантический
поиск по данным

Команда ai в поиске
и правилах

Находит то, что по
словам не ищется



02

Реагирование

Шаблоны действий

Автотриаж на
переходе в работу

Подготовка отчетов



03

AI Observability

Состояние LLM-
сервисов и GPU

Токены, стоимость,
латентность

Трассировки агентов



04

AI Security

Угрозы LLM по
OWASP Top 10

Права агентов и
кража модели

Соответствие ФСТЭК
117

Работа с данными

Раздел 01 · Поиск артефактов атаки там, где
регулярные выражения бесполезны



Эволюция Search Anywhere

ИСТОЧНИКИ

Данные

логи, метрики, трейсы

Документы

регламенты, инструкции

ЯДРО

Smart Monitor

- | search
- | where
- | stats

- | semanticsearch
- | ai

ПРИМЕНЕНИЯ

Аналитика

Правила

Инциденты

Мониторинг

Семантический поиск

AI-обработка

Новые команды Smart Monitor Language

Было и остаётся

source, search, where, stats
timechart, lookup, join
append, eval

И десятки других команд языка
Smart Monitor Language

semanticsearch

Поиск по смыслу, не по словам
Находит близкие формулировки

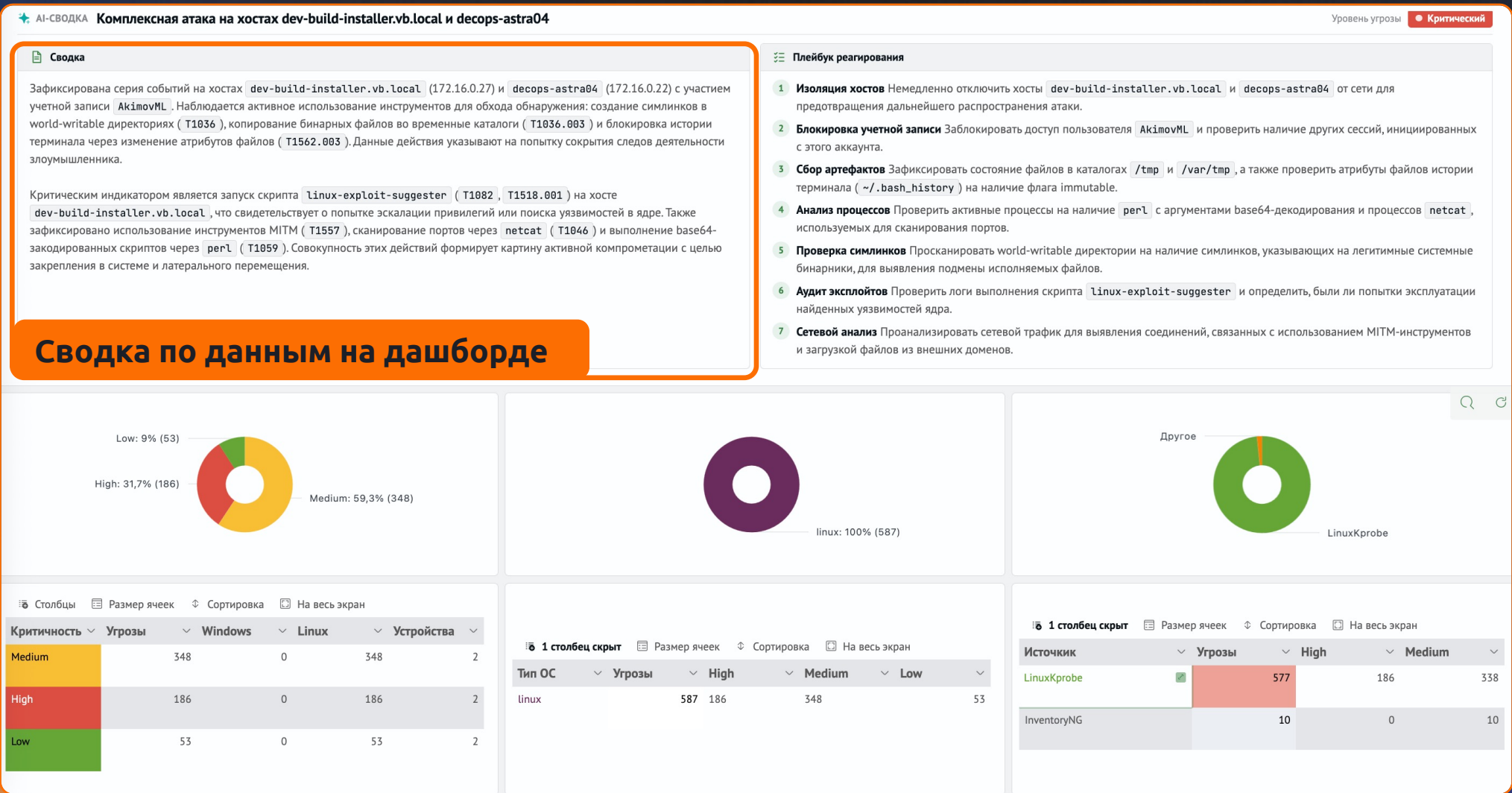
ai

Анализ строк моделью
Результат — в дашборд, правило,
инцидент

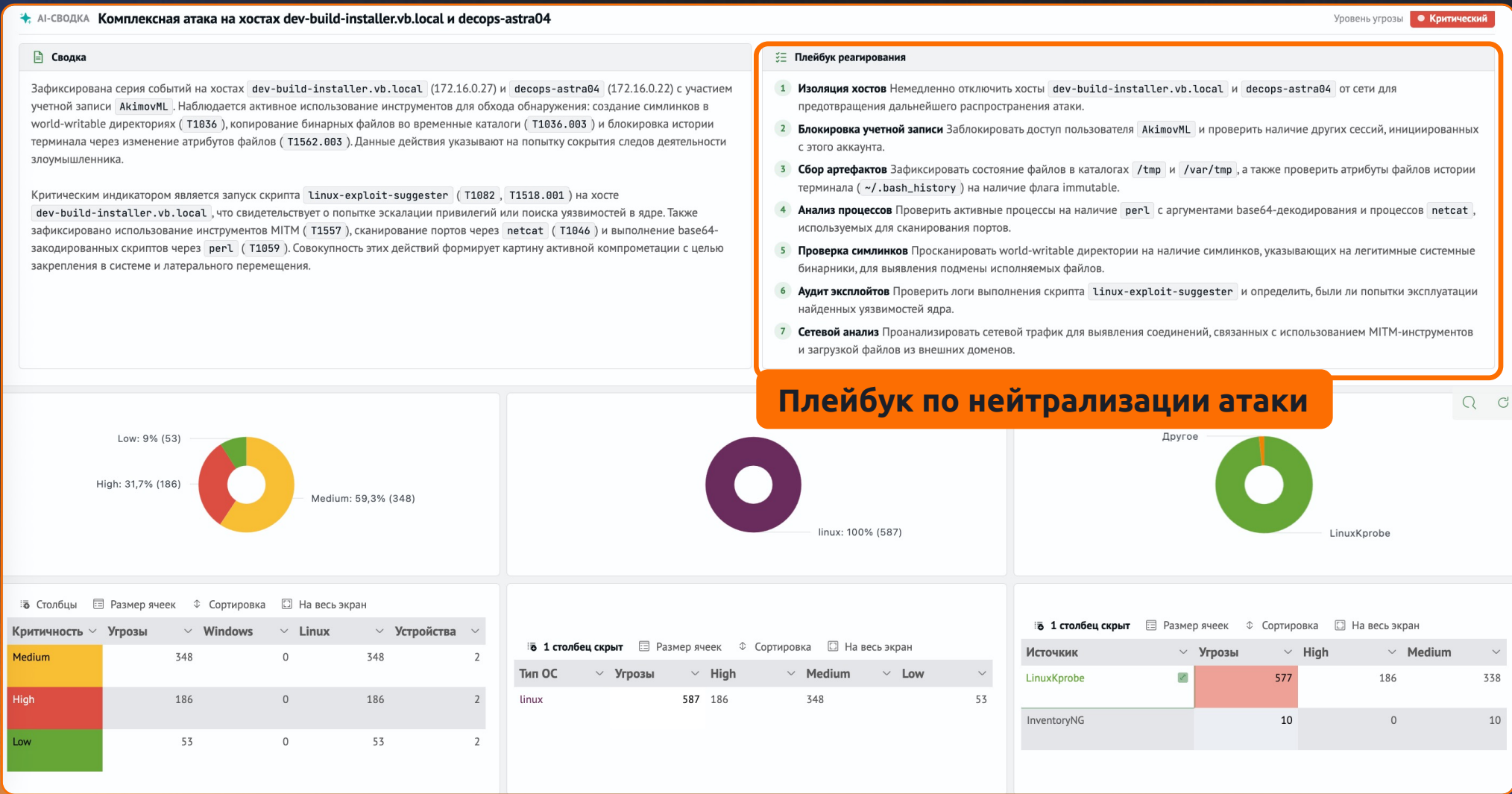
Пример: семантический поиск + инференс модели

```
source message_tracking
| semanticsearch
  query="фишинг и социальная инженерия"
| ai
  user_message="Выявление фишинга и категоризация по критичности"
  output="subject, risk, reason"
```

Сводка по данным на дашборде



Плейбук по нейтрализации атаки



И другие задачи анализа данных в SOC

semanticsearch и **ai** — обычные команды языка запросов



Сводка по инциденту

Резюме из алертов, событий и контекста

```
source siem_incidents  
| ai user_message="напиши резюме по алертам"
```



Разбор аномалий в логах

Похожие аномалии и тип проблемы без regex

```
source app_logs  
| semanticsearch query="похожие аномалии"
```



Триаж потока алертов

Приоритет, тип и обоснование алерта

```
source alert_queue  
| ai user_message="укажи приоритет и тип алерта"
```



Похожие инциденты в истории

Что уже видели и как закрывали

```
source incident_history  
| semanticsearch query="похожие инциденты"
```

Реагирование

Раздел 02 · Активные действия и автотриаж в
Менеджере инцидентов



Первая линия разбирает шум вручную

Ampire 360°

Ложные срабатывания — проблема
детектирования №1



73%

Перегружены объёмом алертов

51%

15–30 мин

ручной разбор одного алерта

свыше 25%

времени аналитика — на ложные
срабатывания

AI помогает не только с данными, но и с процессом

Работа с данными

Сырые данные

Источники и
нормализация

Единое хранилище



Запрос и поиск

Запрос на
естественном языке

Семантический
поиск по событиям



Работа с инцидентом

Инцидент

Триаж и вердикт в
карточке

Проверка версии с
обоснованием

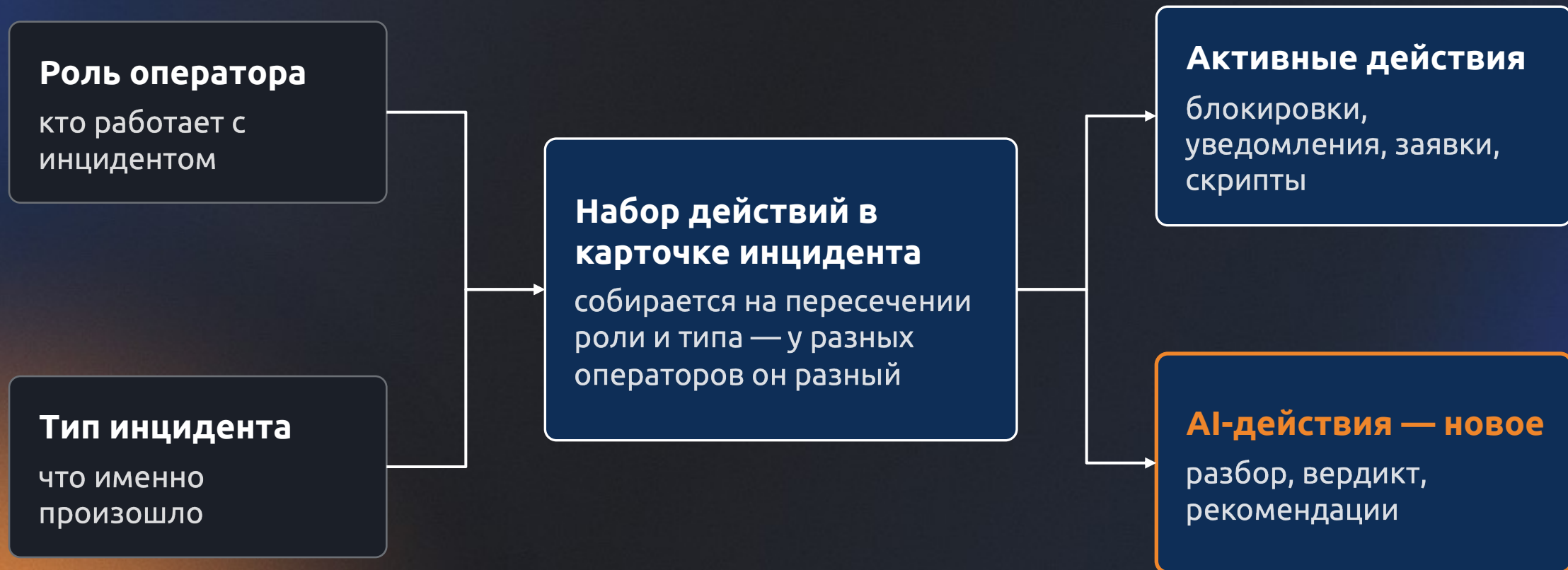


Решение

Шаги реагирования

Итог для смены и
отчёта

Новые AI действия в Smart Monitor



AI действия для задач первой линии

Дата и время	Критичность	Инцидент	Статус
2026-09-22 17:37:01		Пользователь (KrjuchkovaNI) выполнил перебор файлов и каталогов или поиск информации на хосте (VLG-WS363) T1083 Lab	В РАБОТЕ   

[ИИ] Анализ пользователя: KrjuchkovaNI

`date_of_dismissal` = 2021-01-07 — уволена 5 лет назад
`work_status` = Working
`remote_access` = true

4 инцидента за 30 дней — все тактика Discovery

Критический индикатор компрометации: активная учётная запись с удалённым доступом у уволенного сотрудника

Поиск

[ИИ] Обогащение контекстом

[ИИ] Объяснение события

[ИИ] Тriage и приоритет

[ИИ] Проверка версии: TP или FP

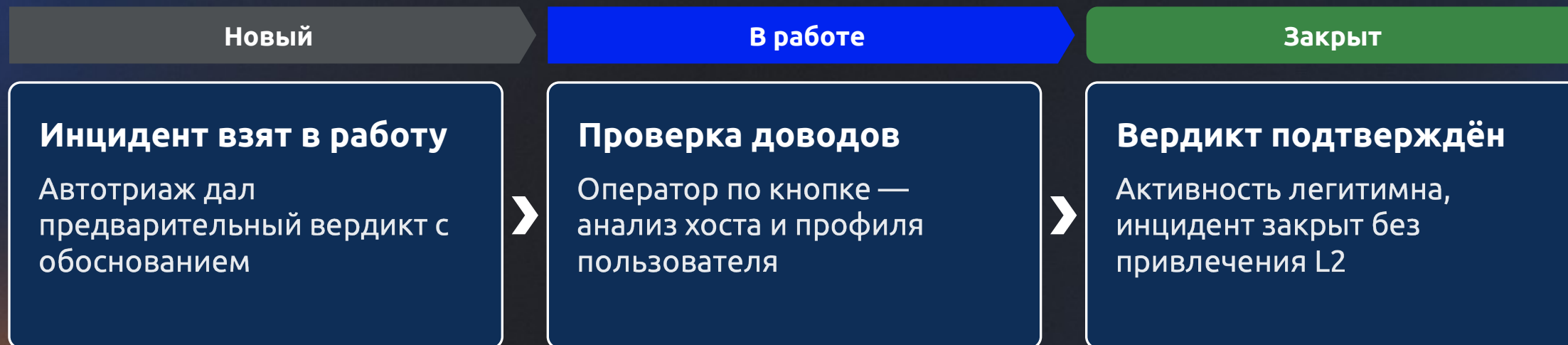
[ИИ] Вердикт и рекомендации

[ИИ] Анализ хоста и пользователя

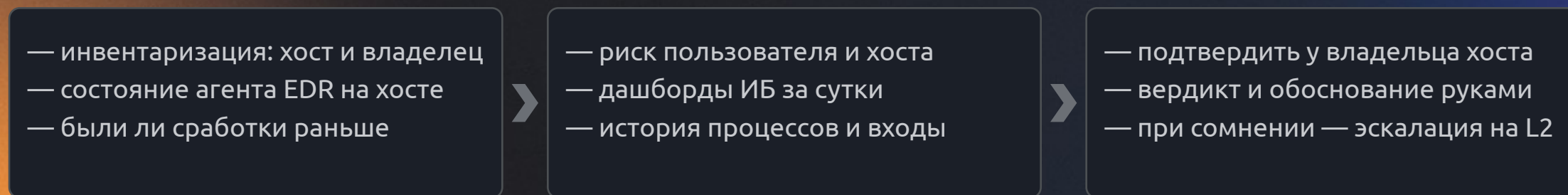
Первая линия закрыла инцидент без эскалации

Пример

Инцидент: попытка доступа к учётным данным на рабочей станции



Без AI то же самое — вручную

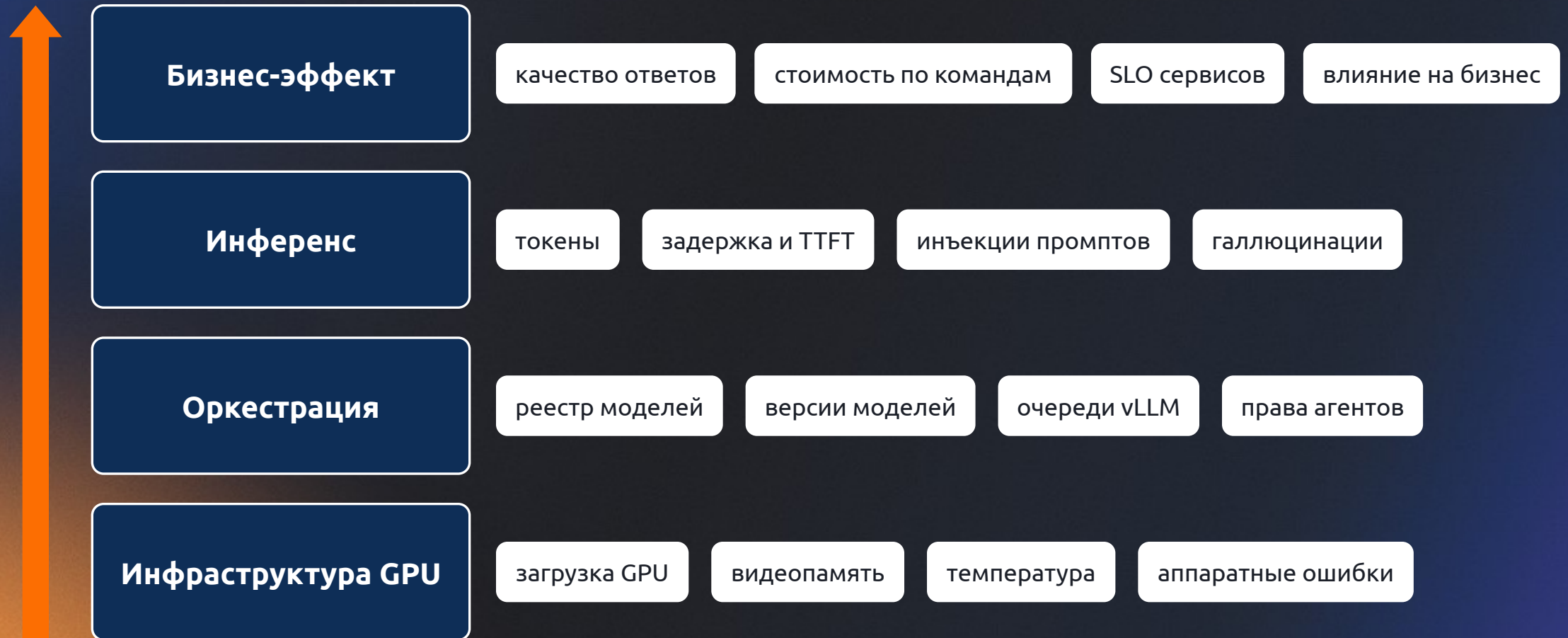


AI Observability

Раздел 03 · Состояние, стоимость и поведение
ИИ — в привычных инструментах SOC



Классический мониторинг не видит ни одного из четырёх слоёв



Источники приведены к одному контракту gen_ai.*



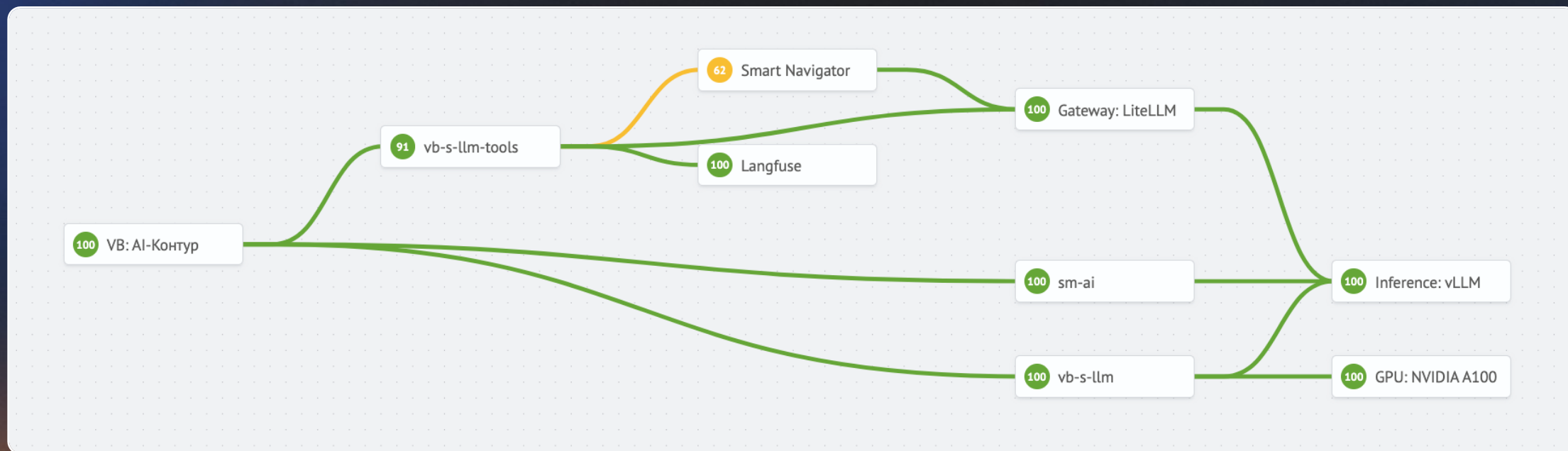
Обращение к LLM даёт множество фактов, которых нет в SIEM

Событие инференса: запрос к vLLM с локального ПК

Кто спросил	Что спросил	Какой сервис	Какая модель	Нагрузка	Какой диалог
a.petrov	«Сравни договор с...»	ai-assistant-support	Qwen3-32B	1286 · 822 ms	516102a3...

```
@timestamp 2026-09-21T11:33:05.000Z  event_type request_completed  host 10.20.0.11
service_name ai-assistant-support  service_type product  service_namespace ai-monitoring
model_name Qwen/Qwen3-32B  model_group vllm/Qwen3-32B
input_tokens 615  output_tokens 671  total_tokens 1286  request_duration_ms 822
session_id 516102a3f57fc4a59af78c81fc7dbbed  user_name a.petrov
prompt_text Сравни договор с клиентом и выдели риски по штрафам...
```

Выявление первопричин деградации через Модель здоровья



Сервис: VB: AI-Контур

**Пользователи
жалуются на
медленные ответы**

Gateway: LiteLLM

**Выросли задержка
ответа и TTFT**

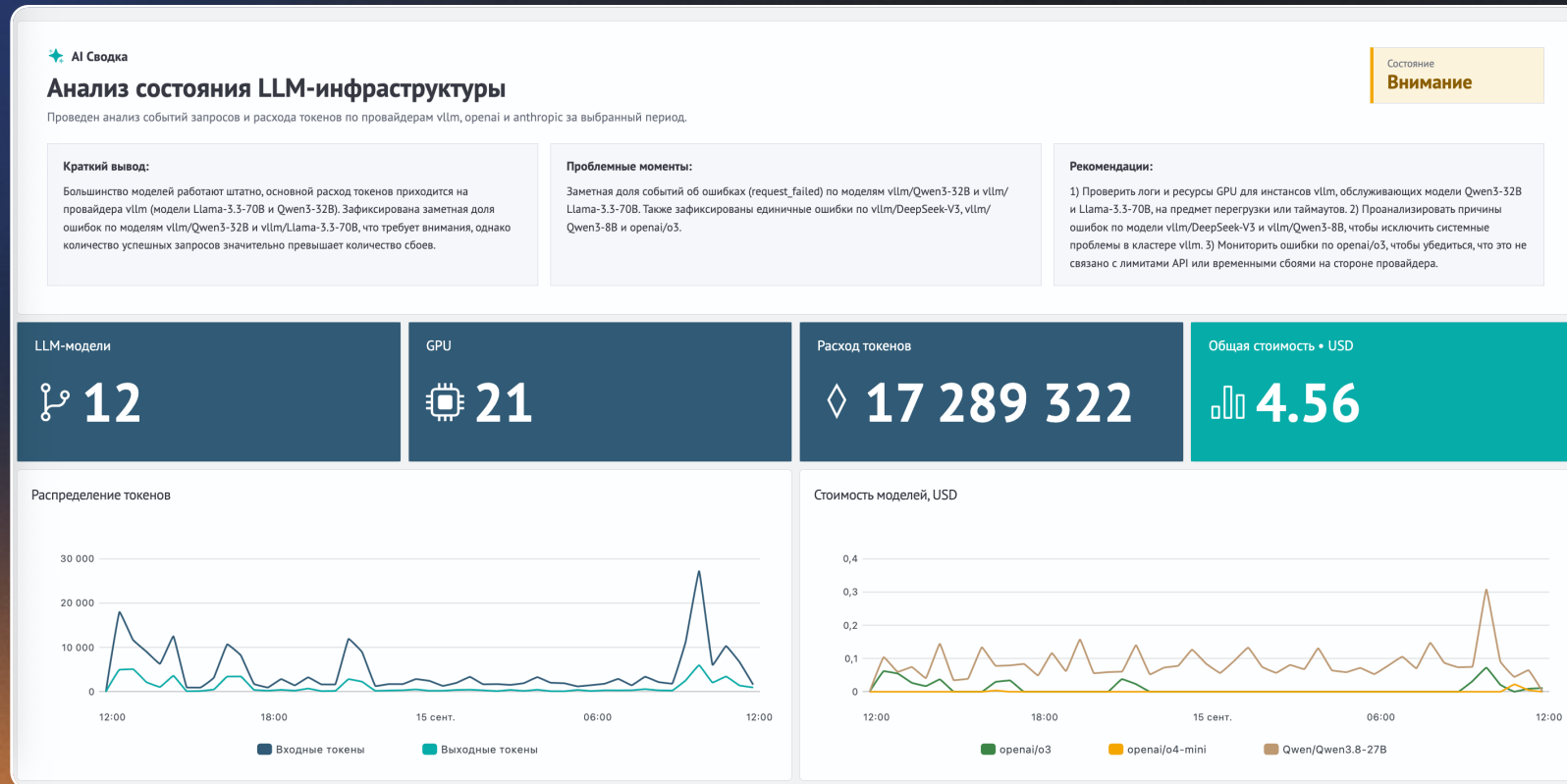
Inference: vLLM

**Очередь в vLLM не
разбирается**

GPU: NVIDIA A100

**Троттлинг по
температуре —
первопричина**

AI Observability: состояние LLM-инфраструктуры в одном окне



Расход и стоимость

по моделям, командам и проектам

Нагрузка и задержки

очередь, TTFT, утилизация GPU

Активность клиентов

кто и с какой машины обращается

Расходы, сбои и качество AI перестают быть сюрпризом

Сегодня		С наблюдаемостью AI-контура
 Расходы	видны в счёте за месяц	считаются в реальном времени по моделям и командам
 Сбои	жалобы есть, диагностики нет	очереди, GPU и задержки лежат рядом с жалобой
 Качество	о деградации сообщают пользователи	видно, что изменилось после смены версии модели
 Теневой AI	не видно совсем	активность внешних клиентов с рабочих машин — в мониторинге

AI Security

Раздел 04 · Выявление атак на LLM-инфраструктуру



Слепая зона современного SOC

Появилось в инфраструктуре

LLM-сервисы и инференс

AI-агенты и MCP-инструменты

GPU, очереди, стоимость запросов

Внешние LLM через API

Claude и Codex на рабочих машинах

SOC видит

Сеть, адреса, порты, протоколы

Известные индикаторы и сигнатуры

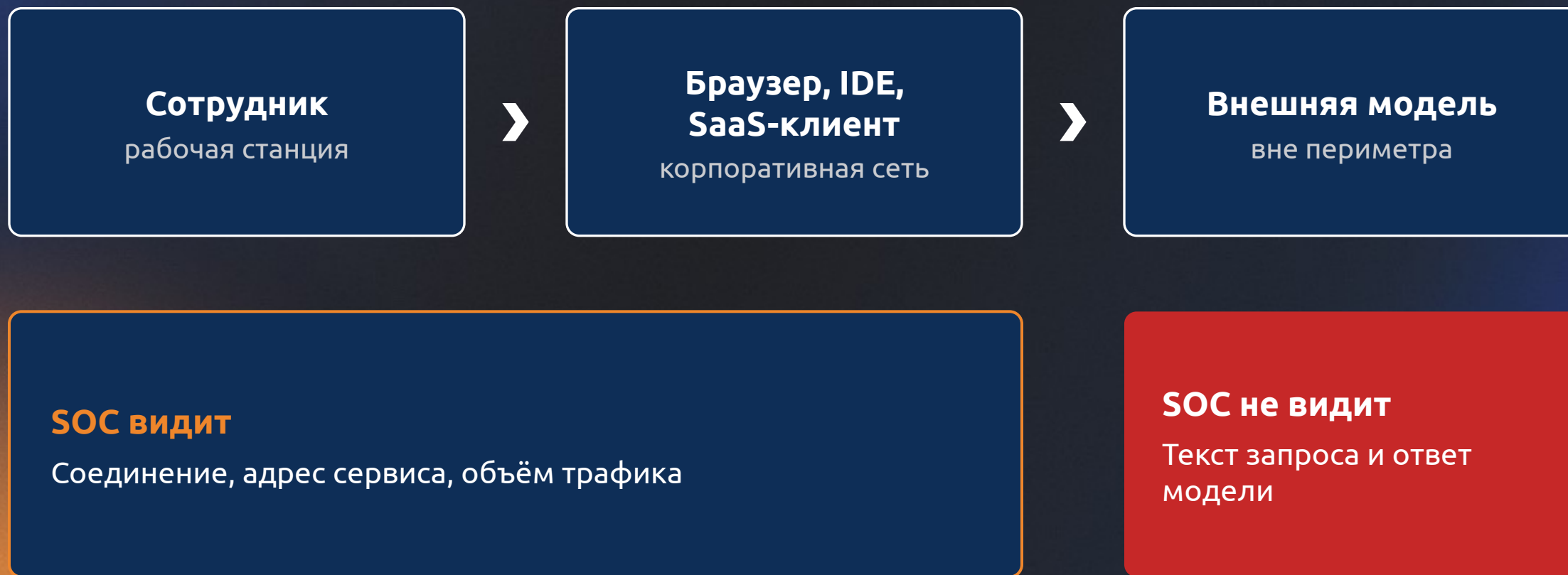
SOC не видит

Тело промпта и ответа модели

Вызовы инструментов и права агента

Расход токенов и утечки в ответах

Слепая зона есть и у тех, у кого своего LLM-контура нет



Покрытие строим по OWASP Top 10 для LLM

Примеры угроз: что именно выявляем в AI-контуре



Инъекции промптов

Прямая и косвенная инъекция в контенте, с привязкой к источнику данных

OWASP LLM01



Избыточные полномочия

Превышение полномочий, злоупотребление инструментами и плагинами агента

OWASP LLM06



Утечка данных и промпта

Раскрытие конфиденциальной информации, эксфильтрация, извлечение системного промпта

OWASP LLM02 / LLM07

Модуль AI Security: Инфраструктура и агенты

Ampire 

Инфраструктура AI-контура
сервисы, шлюзы, MCP



Агент на рабочих станциях
Claude, Codex, ассистенты



Smart Monitor AI Security

сбор и нормализация · правила OWASP · эталонные профили агентов



LLM-ИНФРАСТРУКТУРА

Сводка угроз LLM

счётчики по типам угроз, динамика, активные правила, пользователи и агенты

Инциденты

инъекция промпта, утечка данных, раскрытие системного промпта

АГЕНТЫ И АССИСТЕНТЫ

Безопасность агентов и кража модели

чрезмерные права, нетипичные внешние соединения, отклонения от профиля

Инциденты

превышение полномочий, доступ к артефактам модели

Аналитика по сработкам OWASP Top 10

Ampire 



Картина по типам угроз
счётчики по OWASP Top 10
LLM и динамика
срабатываний

Кто и через что
пользователи, агенты и
сервисы в срабатываниях

Активные правила
какие детекты включены
и как часто срабатывают

Сравниваем каждый вызов агента с его же профилем

Отклонение хотя бы по одному параметру — инцидент OWASP LLM06

Параметр	Эталонный профиль агента	Текущая сессия
 Хост и пользователь	закреплённый хост и учётная запись	совпадает
 Режим запуска	разрешённый режим и sandbox	совпадает
 Инструменты и MCP	разрешённые MCP-серверы и вызовы tools	подключён MCP-сервер вне профиля
 Права и области доступа	минимально необходимые scope	совпадает
 Сеть и origins	разрешённый список origins	совпадает

Без наблюдения за LLM

ЦЕПОЧКА СОБЫТИЙ	01 Письмо со скрытой инструкцией	02 Агент читает вложение	03 Агент собирает данные	04 Ответ модели с данными	05 Отправка на внешний адрес
Почтовый шлюз	видно				
Лог инференса — не собирается					
Сеть, прокси и DLP					видно

Данные ушли: связать события нечем — инцидента нет

Под мониторингом весь контур LLM — цепочка видна целиком

ЦЕПОЧКА СОБЫТИЙ	01 Письмо со скрытой инструкцией	02 Агент читает вложение	03 Агент собирает данные	04 Ответ модели с данными	05 Отправка на внешний адрес
Почтовый шлюз	видно				
Лог инференса с телом запроса и ответа		видно	видно	видно	
Сеть, прокси и DLP					видно

Инцидент зафиксирован: эксфильтрация данных через AI-агента

OWASP LLM02 · ATLAS AML.T0057

Покрытие MITRE ATLAS

ATLAS v5.1 · 12 тактик AI-контура

TA0002 Разведка	TA0003 Ресурсы	TA0000 Доступ	TA0004 Модель	TA0005 Выполнение	TA0006 Закрепление	TA0012 Подготовка	TA0009 Сбор	TA0013 Утечка	TA0010 Воздействие	TA0015 Управление	TA0008 Обход
T0000	T0003	T0051	T0040	T0053	T0020	T0022	T0025	T0026	T0029	T0074	T0015
T0001	T0039	T0010	T0041	T0067	T0043	T0023	T0035	T0057	T0031	T0075	T0050
T0002	T0049	T0052	T0042	T0068	T0054	T0024	T0055	T0073	T0034	T0076	T0030
T0056	T0070	T0065	T0061	T0069	T0071	T0048	T0072	T0058	T0059		T0064
T0044		T0066									
T0060											



закрыто правилами



частично, с обогащением



пока не закрыто

Приказ №117 закрывается тем же контуром наблюдения

Требование приказа №117	Что это значит на практике	Чем закрывается
Логирование запросов и ответов модели	Тело промпта и ответа с метаданными вызова	Слой инференса, единый контракт gen_ai.*
Детектор аномалий, фильтрация входа и выхода	Правила на инъекции, утечки, обход ограничений	Два рубежа: пороги и семантика
Инвентаризация ИИ-активов	Модели, версии, датасеты, промпты — под учётом	Реестр моделей, слой оркестрации
Контроль целостности и доступа	Кто менял промпты, конфигурации и права агентов	Логи платформы и Менеджер инцидентов

Итоги



Пять вопросов, на которые теперь есть ответ

Кто в компании работает с LLM и какими сервисами?



AI Observability

аналитика использования LLM-инфраструктуры

Что именно уходит в модель и что она отвечает?



AI Observability

телеметрия запросов и ответов

Какие из этих запросов — атака?



AI Security

сводка угроз по OWASP и инциденты

Не вышел ли агент за свои полномочия?



AI Security

профиль агента, права и MCP

Как усилить первую линию SOC?



Incident Manager

активные действия, пресеты и автотриаж



Сайт



Telegram



Smart Monitor

powered by  VolgaBlob