SBER TECH | PLATFORM V SBER

Комплексная защита ИИ в СберТехе

От AI Red Team до AI Guardrails



Дмитрий Вандышев

Владелец продукта
Platform V SOWA, СберТех

OWASP Top 10 для LLM

Список ключевых угроз, специфичных для больших языковых моделей

LLM01

Prompt Injection

Злоумышленник вставляет в запрос «секретные команды», и ИИ выполняет их, игнорируя правила

LLM02

Sensitive Information Disclosure

LLM может случайно выдать конфиденциальные данные из обучения или диалога, не отличая их от общедоступных

LLM03

Supply Chain Vulnerabilities

Риски от сторонних компонентов LLM: плагинов, датасетов, предобученных моделей

LLM04

Data and Model Poisoning

Целенаправленное искажение обучающих данных или дообучения для создания бэкдоров, снижения качества или внедрения предвзятости

LLM05

Improper Output Handling

Вывод LLM без проверки передаётся в систему (фронтенд/бэкенд)

LLM06

Excessive Agency

LLM-агентам дают слишком много прав и инструментов без контроля человеком

LLM07

System Prompt Leakage

Атака вынуждает LLM раскрыть системный промпт — правила, конфигурацию, ключи и секреты

LLM08

Vector and Embedding Weaknesses

Атаки на RAG-системы, использующие векторные БД для поиска контекста

LLM09

Misinformation

LLM генерирует правдоподобную, но ложную или предвзятую информацию — не всегда из-за атак, но опасно из-за доверия пользователей

LLM10

Unbounded Consumption

Атака, при которой злоумышленник заставляет LLM-приложение выполнять чрезмерно ресурсоёмкие операции

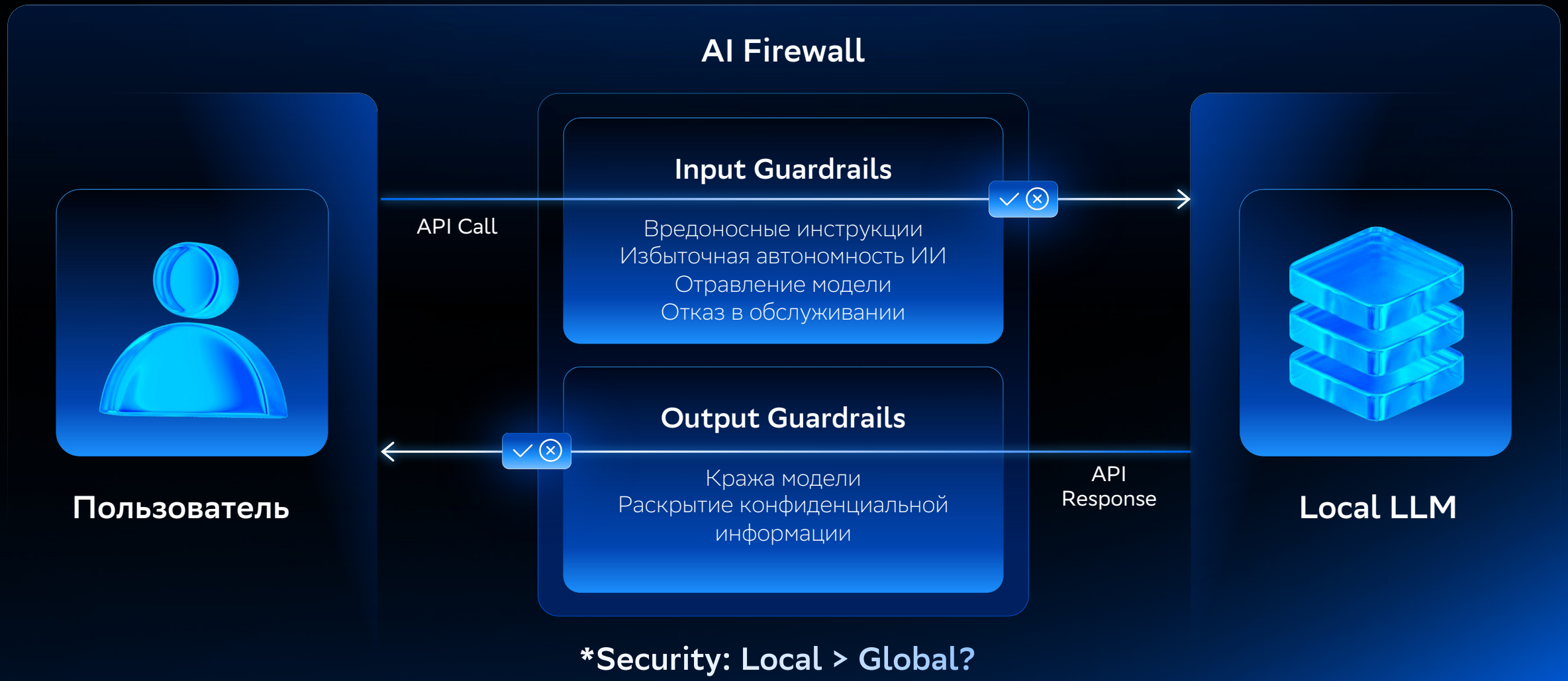
Актуальные кейсы

Кейс 1 Защищаем активы компании при взаимодействии с внешним LLM



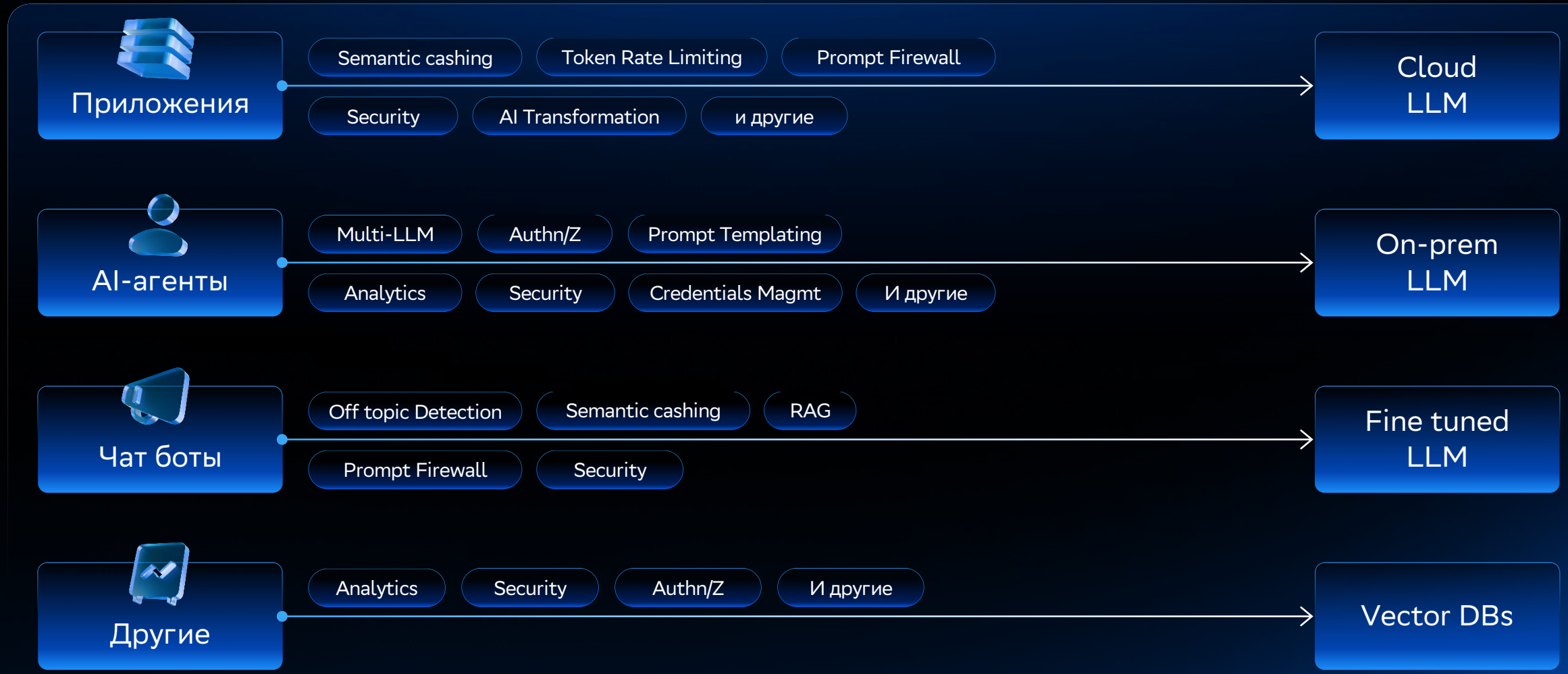
Актуальные кейсы

Кейс 2 Защищаем локальный LLM от внешнего вмешательства



*Security: Local > Global?

Хаотичная архитектура AI-взаимодействий



Как выглядит защищённая экосистема AI-трафика



Инструменты защиты ИИ в СберТехе

Защита ИИ

Platform V SOWA 01

Шлюз для защиты ИИ-систем с встроенными AI-Guardrails

AI Red Team 02

Имитация атак на системы ИИ для выявления уязвимостей и предотвращения рисков

Platform V IAM 03

Авторизация действий AI-агентов



Шлюз безопасности AI

Platform V SOWA

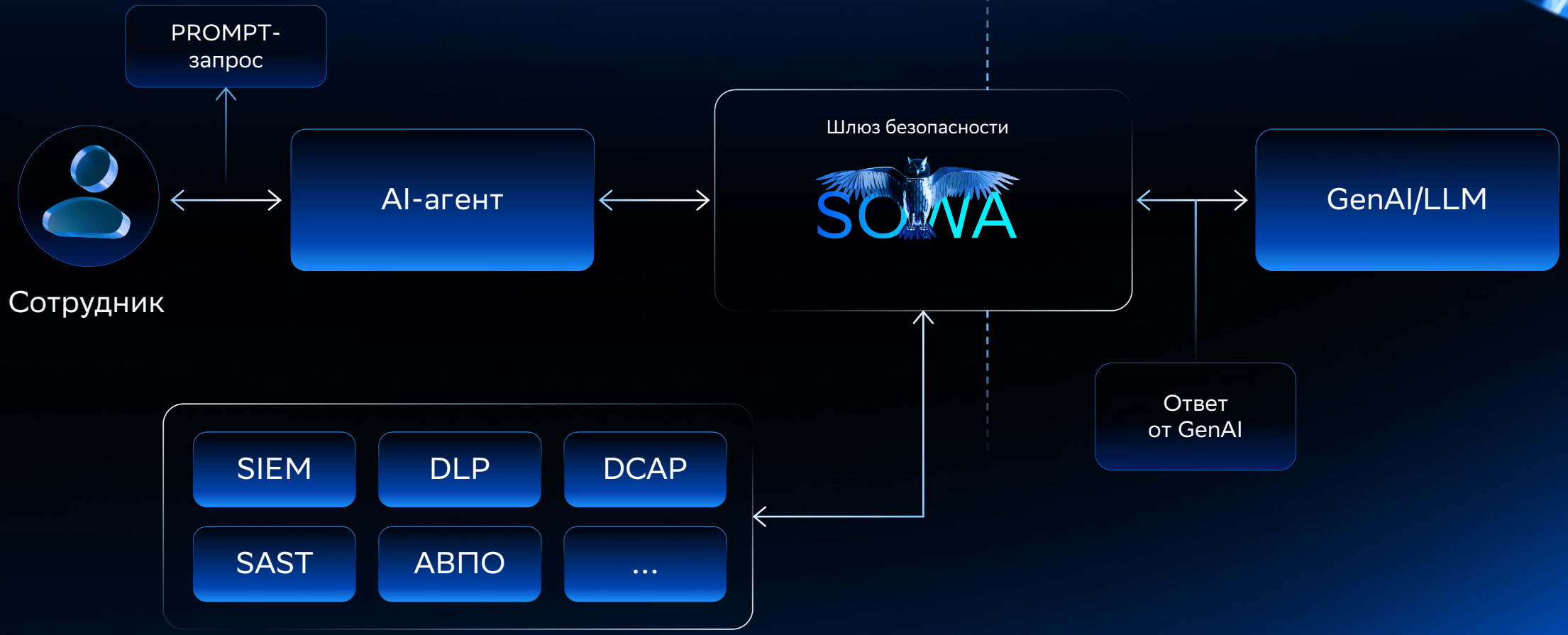
API/AI Security Gateway, который обеспечивает безопасность и интеграцию API в различных сегментах сети и каналах обслуживания

Преимущества для защиты интеграции с GenAI:

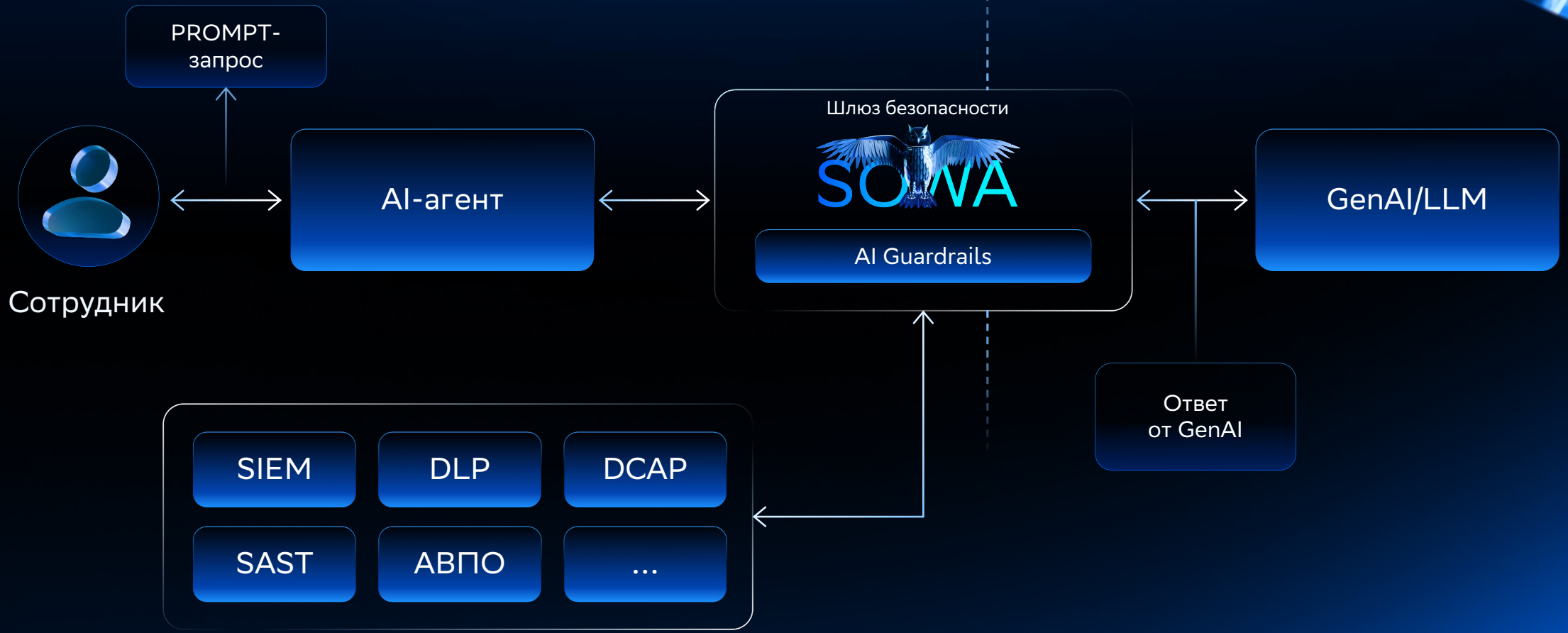
- Ограничение прямого доступа к backend-системам
- Поддержка кастомных протоколов взаимодействия с ИИ
- Автоматическая проверка запросов и ответов для предотвращения утечек данных
- Интеграция с системами DLP, DCAP, АВПО, SAST и др.
- **Наличие AI Guardrails для защиты AI-трафика**



Безопасное взаимодействие с GenAI



Безопасное взаимодействие с GenAI



AI Guardrails: дополнение к Platform V SOWA

AI Guardrails

Это механизмы, которые ограничивают поведение ИИ для обеспечения безопасности, этичности и соответствия заданным нормам



AI Red Team от Лаборатории проверки ПО

Что такое AI Red Team от СберТеха?

Эксперты Лаборатории проверки ПО имитируют действия злоумышленников и атакуют LLM, RAG, плагины и всю GenAI-инфраструктуру так же, как это сделали бы реальные киберпреступники или APT-группы

Цель найти уязвимости до того, как их используют злоумышленники



Ключевые услуги

Глубокое тестирование GenAI-приложений

- Полное покрытие OWASP Top 10 LLM
- Детальный аудит плагинов, RAG-пайплайнов, внешних источников данных
- Четкий отчет с приоритетами и пошаговыми рекомендациями

Эмуляция атак реальных APT-групп на AI-инфраструктуру

- Индивидуальные сценарии под ваш AI-стек
- Симуляция цепочек атак
- Проверка на отказ в обслуживании, манипуляцию выводов, обход AI Guardrails

50+

LLM и AI-агентов протестировано

120+

Уязвимостей выявлено в AI-системах

Platform V IAM

Сертифицированное ФСТЭК решение на базе Keycloak, обеспечивающее enterprise-уровень безопасности, производительности и масштабируемости

Наши преимущества:

Использование известного и понятного open-source движка

Гарантирует профессиональную и легальную поддержку

Безопасность с сертификацией ФСТЭК

Быстрое устранение уязвимостей, наличие сертификации для строгих требований

Высокая производительность

Оптимизирован для высоких нагрузок и отказоустойчив

Enterprise-масштабирование

Гибкая кластеризация и распределение нагрузки, поддержка enterprise ролевых моделей и протоколов

Нагрузка и масштабируемость:

400+

тысяч пользователей в каталоге AD

на 3 ЦОД

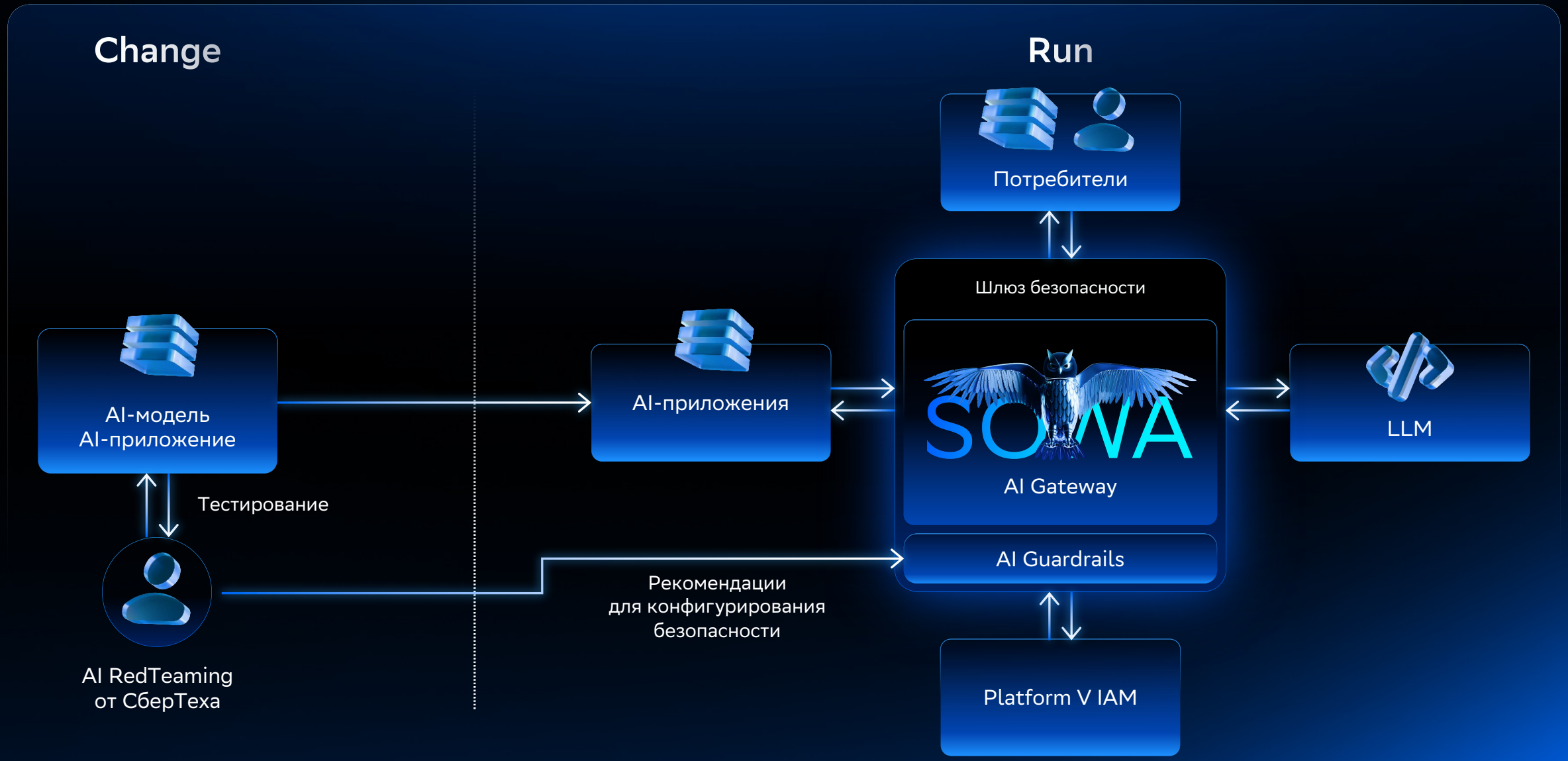
готовность к развертыванию

2000+

инсталляций в Сбере

- Оперативная поддержка от команды экспертов СберТеха
- Актуальная документация, гайды и кейсы ускоряют обучение и онбординг
- Регулярные проверки и патчи от СберТеха для минимизации уязвимостей

Комплексная защита OWASP Top 10 LLM



SBER TECH | PLATFORM V SBER

Спасибо за внимание!

Контакты

Почта: vandyshev.d.v@sbertech.ru

TG: @SeaYacht

