ORION SOFT

STARGUARD AI

Практическая защита ML-моделей
и ИИ-решений от киберугроз



ORION SOFT

Orion soft создает защищенное и стабильное ПО для ИТ-инфраструктуры Enterprise-бизнеса.

850+
заказчиков

1000+
авторизованных партнеров

350+
экспертов в команде

8+
лет разрабатываем ИТ-продукты

№1
в рейтинге «Стабильная виртуальная инфраструктура»*

топ-3
в рейтинге «Российские платформы виртуализации»*

№1
в рейтинге лучших ИТ-работодателей России**

топ-3
в рейтинге «Российские платформы Kubernetes»*

** CNewsMarket, 2025*
*** Хабр Карьера, 2025*

Экосистема Orion soft



Как крупные компании развивают использование ИИ

95% из топ-100 крупнейших компаний уже внедрили ИИ-решения *

Сценарии использования

- Обучение сотрудников
- Транскрибация встреч
- ИИ-помощник для рутинных задач
- Применении ИИ в производственных процессах
- «Умный» поиск по базе знаний

* Forbes

Риски массового использования ИИ

Конфиденциальные данные могут попасть в облачные модели

ИИ-помощник может отвечать на вредоносные запросы

ИИ-агенты могут выполнять неразрешенные действия

Сотрудники бесконтрольно расходуют токены

Пользователи получают закрытую для них информацию через RAG-системы, документы и вложения

//Компании выбирают между двумя подходами:

- **Технический запрет:** блокируют использование ИИ на всех корпоративных устройствах
- **Контролируемое использование:** объясняют сотрудникам риски передачи данных через личные аккаунты и предоставляют удобные корпоративные системы



STARGUARD AI

ЗАЩИТА ВЗАИМОДЕЙСТВИЯ С LLM

Платформа для безопасного
и контролируемого взаимодействия
с большими языковыми моделями
(Large Language Models, LLM)

- Шлюз между конечными пользователями, прикладными системами и LLM
- Система для комплексного анализа запросов и ответов из LLM в режиме реального времени



Основные функции StarGuard Gate

Детектор угроз и утечек данных

//Проверка по регулярным выражениям

Сопоставляет запросы с regex-шаблонами для выявления идентификаторов, кодов, номеров документов

//Семантическая фильтрация

Проверяет запросы и ответы на соответствие политикам; выявляет токсичный и нежелательный контент

///Обнаружение и маскирование данных

Находит и маскирует персональные и конфиденциальные данные до передачи в LLM, восстанавливает их в ответе

//Защита от Prompt Injection

Обнаруживает попытки обхода политик и изменения поведения модели.

//Контроль языка запросов

Определяет язык сообщения и применяет списки разрешенных и запрещенных языков до передачи запроса выбранной модели

//Контроль вызовов инструментов

Проверяет генерируемые моделью вызовы инструментов по отдельным политикам с режимами разрешающего и запрещающего списков

//Проверка файловых вложений

Извлекает текст из DOC, DOCX, XLS, XLSX и PDF и применяет к нему те же детекторы, что и к обычным сообщениям

//Режимы Forward / Reverse proxy

Работает как reverse proxy для прямого доступа через шлюз и как forward proxy для прозрачной проверки обращений к внешним LLM-провайдерам

Основные функции StarGuard Base

Управление доступом и контроль использования

//Авторизация и аутентификация

Поддерживает OIDC/SSO, статический Bearer-токен и режим без аутентификации в зависимости от сценария подключения

//Наблюдаемость

Предоставляет журнал событий, трассировку запросов, метрики использования и данные для интеграции с системами мониторинга и SIEM

///Лимиты потребления токенов

Учитывает расход токенов по каждому пользователю и автоматически блокирует новые запросы после достижения суточного лимита

//Аудит событий безопасности

Сохраняет метаданные запросов, результаты срабатывания детекторов и трассировку; поддерживает передачу событий в SIEM

//Поддерживаемые API-форматы

Поддерживает OpenAI Chat Completions, Completions, Responses API, Anthropic Messages API, а также потоковый режим SSE

//Подключение LLM-провайдеров

Предоставляет единый OpenAI- и Anthropic-совместимый API для работы с внешними и локальными моделями, включая потоковые ответы

//Разграничение доступа

Назначает пользователям и группам права доступа к шлюзам и моделям; поддерживает аутентификацию через корпоративный OIDC

//Управление политиками

Централизованная настройка политик для запросов, ответов и инструментов с настройкой действий детекторов

Кейс №1:
Защита данных при использовании облачных и локальных моделей

/Задача

Предотвратить передачу конфиденциальных данных при работе с облачным ИИ.

//Решение

- Перенаправление запросов с критичными данными в локальную модель
- Маскирование персональных данных для менее критичных сценариев

95%
точность перенаправления

92%
точность маскирования



Кейс №2:
Защита закрытых документов в RAG-системе

/Задача

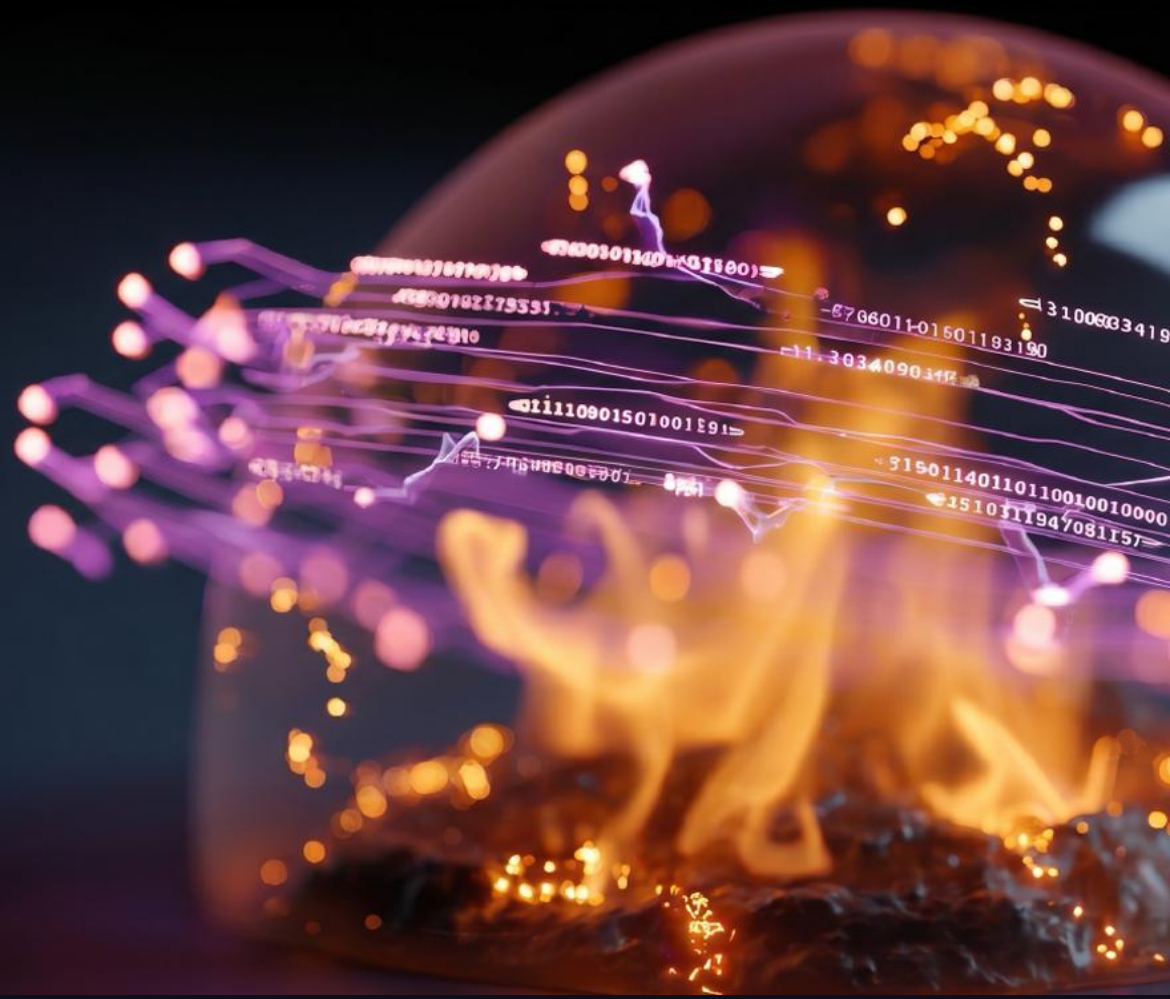
Не допустить, чтобы ИИ-помощник выдал пользователю закрытые сведения.

//Решение

- Проверка прав пользователя при обращении к RAG-системе
- Блокировка утечек конфиденциальной информации из внутренней базы знаний

97%

точность по заблокированным запросам



Кейс №3:
Защита взаимодействия ИИ-агентов

/Задача

Обеспечить контроль над мультиагентной системой.

//Решение

- Создание «белых списков» разрешенных инструментов для агентов
- Следующий этап: дополнение политик доступа и проверка в интеграционных сценариях

82%

точность на тестовых данных



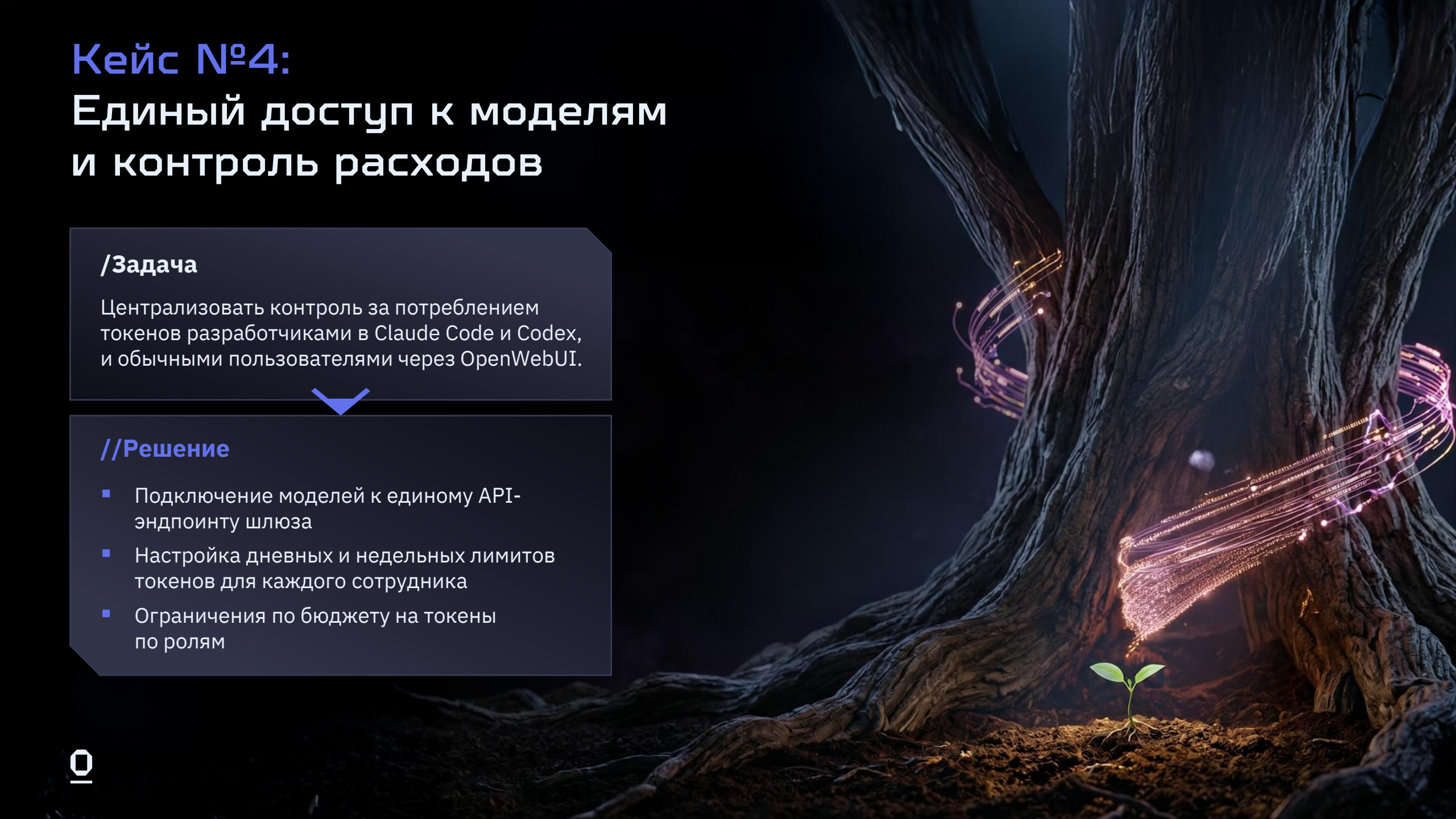
Кейс №4:
Единый доступ к моделям и контроль расходов

/Задача

Централизовать контроль за потреблением токенов разработчиками в Claude Code и Codex, и обычными пользователями через OpenWebUI.

//Решение

- Подключение моделей к единому API-эндпоинту шлюза
- Настройка дневных и недельных лимитов токенов для каждого сотрудника
- Ограничения по бюджету на токены по ролям



Архитектура StarGuard AI

Прикладные системы

- OpenWebUI
- Корпоративный чат-бот
- Внутренние ИИ-сервисы
- Бизнес-приложения
- ИИ-агенты
- Другие решения

Запрос

Маскирование данных
Подмена ПДн спецтокенами (опционально)

Regex detector
Поиск заданных шаблонов в запросе

NER detector
Распознавание ПДн и других сущностей

LLM detector
Смысловая проверка запроса и контекста

Принятие решения
Пропуск, блокировка, добавление сообщения, перенаправление



LLM
Облачные и локальные модели

LLM detector
Смысловая проверка ответа модели

Контроль Tool Call
Проверка и фильтрация вызовов инструментов (whitelist/blacklist)

Демаскирование
Восстановление данных в ответе (опционально)

Ответ

---------- Опциональный элемент

* Набор детекторов может быть расширен

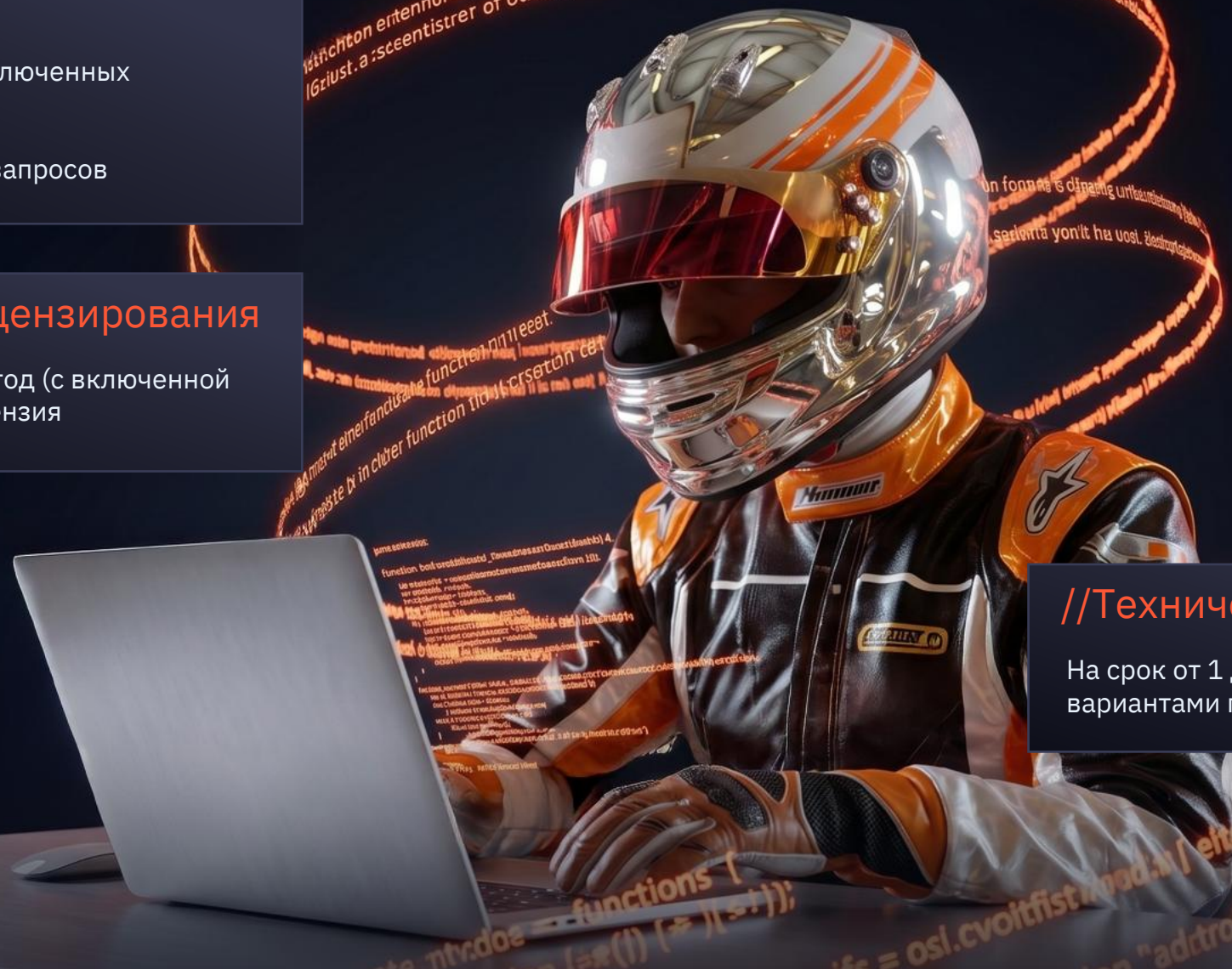
Лицензирование и ТП

//Лицензии

- По количеству подключенных пользователей
- По числу суточных запросов

//Варианты лицензирования

Срочная лицензия на год (с включенной ТП) и бессрочная лицензия



//Техническая поддержка

На срок от 1 до 5 лет с разными вариантами по SLA (9/5 или 24/7)

Также в экосистеме Orion soft:

NOVA AI

Первая российская Kubernetes-платформа, оптимизированная под задачи ИИ и машинного обучения.

70%

Загрузка GPU вместо типичных 10-30%

1 день

До готовой инфраструктуры для ИИ/МО

Какие задачи решает:

Работа с GPU

- Низкая утилизация видеокарт
- Конфликты с использованием общих карт на несколько команд
- Низкая observability

Вывод в prod LLM-сервиса

- Затраты ресурсов на ручную настройку сред
- Нет поддержки платформы 24/7
- Нет ответственных за ИБ

Команда и SLA

- Не хватает квалифицированных специалистов в штате
- Сложный поиск и удержание ключевых сотрудников
- Bottle neck в ключевых людей

Защищенный доступ к AI-моделям без затрат на собственную GPU-инфраструктуру

K2 CLOUD x STARGUARD AI

Актуально

- Компаниям, которым требуется контролируемый и защищенный доступ к облачным AI-моделям
- Командам, у которых нет собственной доступной GPU-инфраструктуры
- Бизнесу, который хочет протестировать решение перед масштабным внедрением

Преимущества

- Гарантированный SLA для всей системы
- Гибкая помесячная подписка
- GPU-ресурсы без капитальных затрат на оборудование
- Быстрый запуск без длительного развертывания инфраструктуры

ORION SOFT

Хотите бесплатно протестировать наши решения? Оставьте заявку!

info@orionsoft.ru

@orionsoftru

Стенд на 90 дней по запросу:

На инфраструктуре вендора / дистрибутора

В вашем периметре

