Innostage
центр
разработок

ОПИСАНИЕ СИСТЕМЫ, ФУНКЦИОНАЛЬНЫХ ХАРАКТЕРИСТИК И ЭТАПОВ ЖИЗНЕННОГО ЦИКЛА

ПРОГРАММНЫЙ ПРОДУКТ

«Защита ИИ. AIDR»

версия 1.0

2026 г.

Оглавление

1. Общие сведения

1.1.Наименование программного продукта:

«Защита ИИ. AIDR».

1.2.Назначение:

Программный продукт «Защита ИИ. AIDR» (далее – Программный продукт, Система) предназначен для автоматизации контроля входных и выходных данных систем искусственного интеллекта (ИИ), обеспечивая безопасность взаимодействия с языковыми моделями (LLM) через валидацию промптов и ответов, фильтрацию нежелательного контента, детекцию персональных данных (PII) и секретов, а также интеграцию с системами мониторинга информационной безопасности (SIEM/SOC).

Настоящий документ описывает функциональные характеристики Программного продукта, а также процессы, обеспечивающие сопровождение и развитие на всех этапах его жизненного цикла, включая его разработку, сопровождение, устранение неисправностей, совершенствование и техническую поддержку.

Документ предназначен для специалистов, обеспечивающих эксплуатацию, сопровождение, установку и интеграцию Программного продукта.

Процессы реализованы в соответствии с требованиями ГОСТ 19.101-77, ГОСТ 19.104-78, ГОСТ Р ИСО/МЭК 12207-2010, международному стандарту ISO/IEC/IEEE ISO/IEC/IEEE 26511:2018 «Systems and software engineering – Requirements for software user documentation» и адаптированы под архитектуру и технологический стек продукта.

Документ содержит сведения об используемых инструментах, персонале, режиме поддержки, применяемых при сопровождении ПО на территории Российской Федерации.

1.3.Область применения:

- Организации, использующие корпоративные LLM-системы.
- Центры обработки данных и облачные платформы.
- Команды информационной безопасности (SOC).
- Государственные информационные системы, обрабатывающие данные с использованием ИИ.

2. Обозначения и сокращения

В настоящем документе используются следующие обозначения и сокращения:

Таблица 1

ОБОЗНАЧЕНИЕ	РАСШИФРОВКА ОБОЗНАЧЕНИЯ
База данных (БД)	Набор данных, который достаточен для установленной цели и представлен на машинном носителе в виде, позволяющем осуществлять автоматизированную переработку содержащейся в нем информации
Вендор	Компания, которая разрабатывает, производит или поставляет ИТ-продукты, оборудование или программное обеспечение под собственным брендом и продаёт лицензии на их использование
Информационная безопасность (ИБ)	Состояние защищенности информации Учреждения от внутренних и внешних информационных угроз, при котором обеспечивается реализация прав работников,

ОБОЗНАЧЕНИЕ	РАСШИФРОВКА ОБОЗНАЧЕНИЯ
	достижение уставных целей, функционирование и устойчивое социально-экономическое развитие [ГОСТ Р ИСО/МЭК 27005-2010]
Интерфейс прикладного программирования (API)	Набор правил и протоколов, который позволяет различным программам взаимодействовать друг с другом и обмениваться данными
ИИ / AI	Искусственный интеллект / Artificial Intelligence
Инференс	Процесс применения обученной модели машинного обучения к новым данным для получения предсказаний или принятия решений, без этапа дообучения модели. Ключевой этап эксплуатации ИИ-моделей в продакшене.
ИТ-актив	Сетевой ресурс
ПО	Программное обеспечение
ОС	Операционная система
СУБД	Система управления базами данных
CPU	Центральный процессор, выполняющий последовательные вычисления и управляющий общими задачами системы
GPU	Графический процессор, оптимизированный для параллельной обработки больших объёмов данных (графика, машинное обучение, научные расчёты)
LLM	Большая языковая модель – тип искусственного интеллекта, основанный на нейронной сети с миллиардами параметров, обученный на огромных объёмах данных
ML	Машинное обучение – раздел искусственного интеллекта, позволяющее компьютерам учиться без прямого программирования, используя статистические алгоритмы для анализа данных
On-prem	ИТ-инфраструктура (серверы, ПО, сети), развёрнутая и управляемая внутри собственных помещений организации
Private Cloud	Выделенная облачная среда, предназначенная исключительно для одной организации, сочетающая преимущества облачных технологий с контролем и безопасностью локальной инфраструктуры
Security information and event management (SIEM)	Класс программных продуктов, предназначенных для сбора и анализа информации о событиях безопасности.
Security Operations Center (SOC)	Централизованное подразделение или сервис, отвечающий за непрерывный мониторинг ИТ-инфраструктуры, обнаружение кибератак и реагирование на инциденты информационной безопасности
Transport Layer Security (TLS)	Криптографический протокол, который обеспечивает безопасную передачу данных между клиентом и сервером в сети Интернет
User Interface (UI)	Часть программы, сайта или приложения, с которой непосредственно взаимодействует пользователь

3. Архитектура продукта

Схема решения Программного продукта представлена ниже (Рисунок 1).

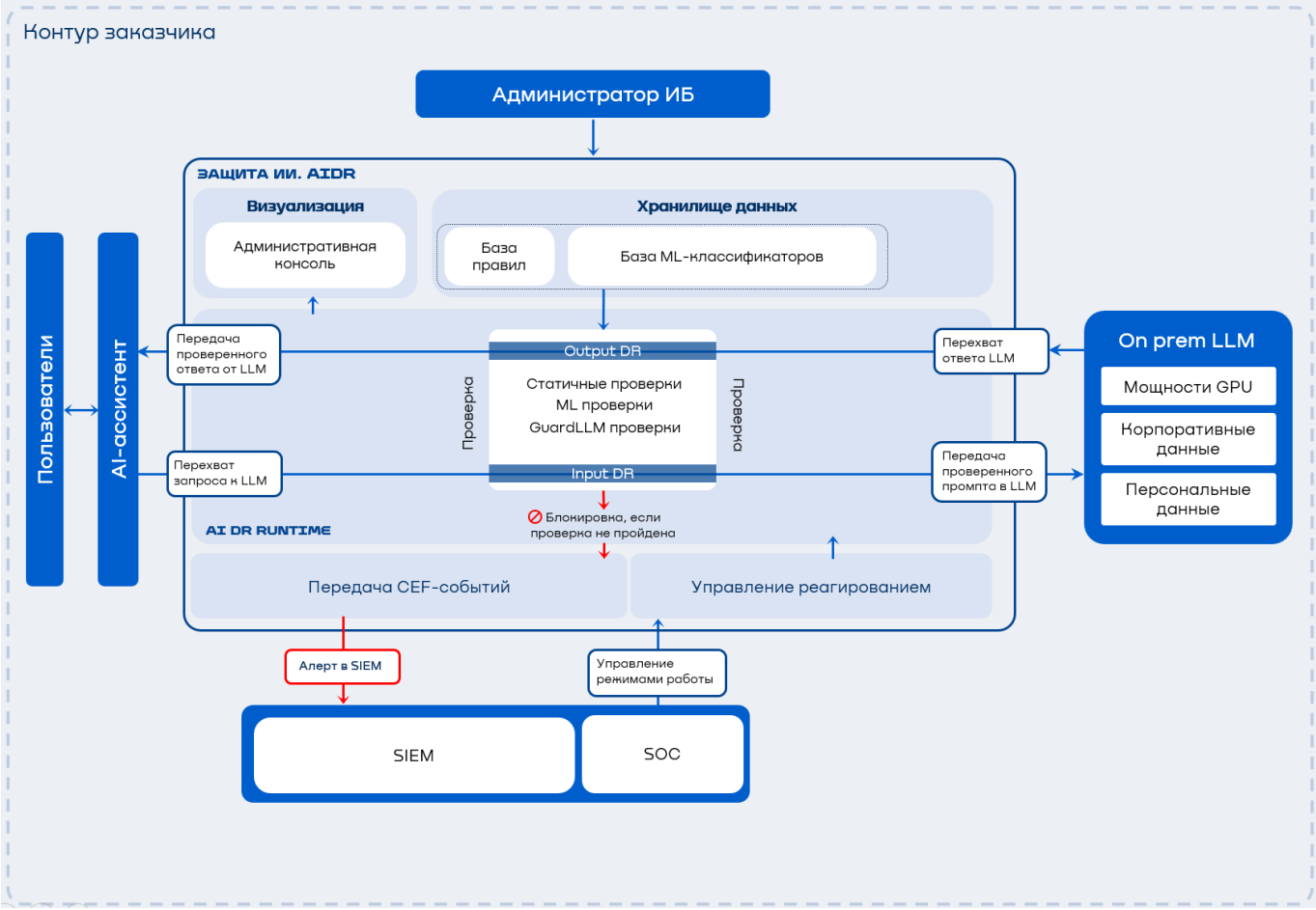

Рисунок 1

3.1.Тип архитектуры:

Микросервисная архитектура с модулями:

- AI Detection and Response Proxy – основной модуль прокси и оркестрации валидаторов.
- static_ag_kt_service - модуль валидации уровня статических проверок на контроль ПДн
- static_lg_service - модуль валидации уровня статических проверок
- ml_grai_service - модуль валидации уровня ML проверок
- Admin Console UI - Административная веб-консоль (Typescript, Vue.js 3).

Все основные компоненты Программного продукта «Защита ИИ. AIDR» используют лицензии Apache-2.0, MIT, BSD-3-Clause, которые разрешают коммерческое использование, модификацию и распространение.

3.2.Контур развёртывания:

Продукт развёртывается в контуре заказчика (On-prem или Private Cloud):

- Модули валидации работают как прокси-шлюз между клиентским приложением и LLM.
- Хранилище данных (правила, ML-классификаторы) размещается в инфраструктуре заказчика.

3.3.Интерфейсы взаимодействия:

- REST API, совместимый с форматом OpenAI Chat Completions.
- Веб-интерфейс администратора (Typescript, Vue.js 3).
- Интеграция с SIEM через Syslog / HTTP Webhook (формат CEF).

4. Функциональные возможности

4.1.Валидация входящих запросов (промптов):

Таблица 2

Функция валидации	Описание требования	Метод реализации	Используемые компоненты
Детекция персональных данных (PII)	Поиск ФИО, адресов, паспортных данных, контактов в запросах пользователей	Регулярные выражения (Regex) + NER-модели	presidio_analyzer, spacy, regex, angryscan-core, llm(guard)
Детекция секретов (secrets)	Обнаружение API-ключей, паролей, токенов, учётных данных	Паттерн-матчинг + Регулярные выражения (Regex) + ML-классификатор на базе RuBERT моделей	detect-secrets, natasha, regex, llm(guard)
Фильтрация небезопасного контента (NSFW)	Блокировка контента 18+, насилия, оружия, экстремизма	ML-классификатор на базе предобученных моделей	transformers, llm(guard)
Контроль темы (Topic control)	Проверка соответствия промпта разрешённому списку тем и списку запрещенных тем	Семантический анализ через векторные эмбеддинги (Embeddings)	transformers, spacy, llm(guard)
Обнаружение Jailbreak / Prompt Injection	Обнаружение попыток взлома системного промпта («DAN», «Ignore previous instructions»)	Классификатор аномальных паттернов + правила корреляции	transformers, llm(guard)

4.2.Валидация исходящих ответов (ответов LLM):

Таблица 3

Функция валидации	Описание требования	Метод реализации	Используемые компоненты
Детекция персональных данных (PII) в ответах	Поиск ФИО, адресов, паспортных данных, контактов в запросах пользователей	Регулярные выражения (Regex) + NER-модели	presidio_analyzer, spacy, regex, angryscan-core, llm(guard)
Детекция секретов в ответах	Обнаружение API-ключей, паролей, токенов, учётных данных	Паттерн-матчинг + Регулярные выражения (Regex) + ML-классификатор на базе RuBERT моделей	detect-secrets, natasha, regex, llm(guard)
Фильтрация NSFW в ответах	Блокировка контента 18+, насилия, оружия, экстремизма	ML-классификатор на базе предобученных моделей	transformers, llm(guard)
Контроль темы (Topic control)	Проверка соответствия промпта разрешённому списку тем и списку запрещенных тем	Семантический анализ через векторные эмбеддинги (Embeddings)	transformers, spacy, llm(guard)
Токсичный контент (Toxic/Obscene)	Обнаружение попыток взлома системного промпта («DAN», «Ignore previous instructions»)	Классификатор аномальных паттернов + правила корреляции	transformers, llm(guard)
Обнаружение Jailbreak / Prompt Injection	Поиск ФИО, адресов, паспортных данных, контактов в запросах пользователей	Регулярные выражения (Regex) + NER-модели	presidio_analyzer, spacy, regex, angryscan-core, llm(guard)

4.3.Административная веб-консоль:

Функционал предназначен для мониторинга и управления Системой.

Таблица 4

Модуль консоли	Функциональность	Технологическая реализация	Используемые компоненты
Конфигуратор валидаторов	Toggle вкл/выкл проверок, настройка порогов чувствительности (0.0–1.0), управление «белыми»/«чёрными» списками	Динамическая валидация конфигов	Pydantic, PyYAML, fastapi
Управление ключами и Endpoint-ами	Настройка подключения к внешним LLM (vLLM, OpenAI-совместимые), настройка API-ключа для доступа приложения	Безопасное хранение секретов	pydantic-settings, python-dotenv, cryptography

Для входа в Систему необходимо выполнить следующие действия:

- ввести учетные данные (логин и пароль) в окне авторизации (Рисунок 2);
- нажать кнопку «Войти в систему».

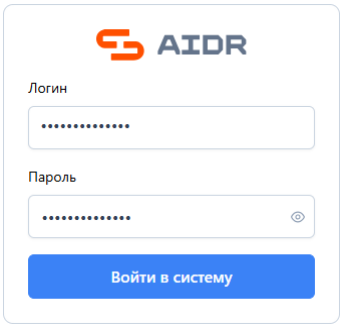

Рисунок 2

После авторизации в Системе пользователю автоматически открывается меню с виджетами функциональных настроек применяемых политик и наборов правил (Рисунок 3).

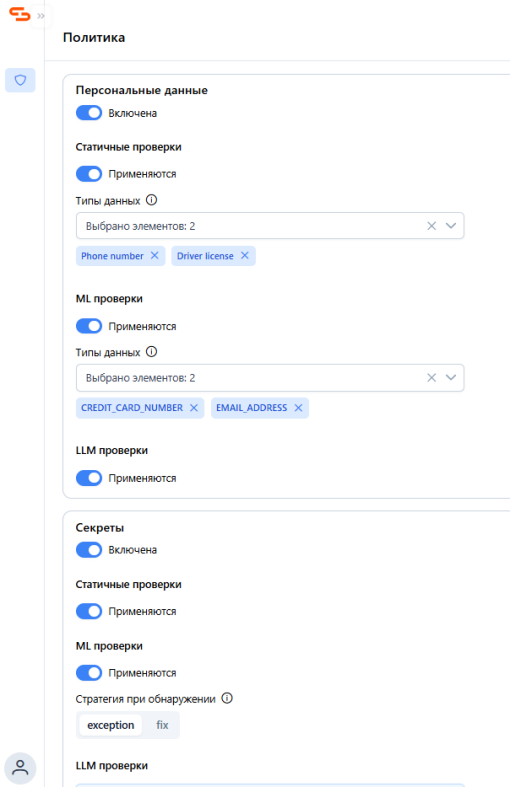

Рисунок 3

Каждый блок с виджетом является редактируемым и позволяет персонализировать настройки политик под конкретные нужды и задачи.

Для установки собственных параметров необходимо воспользоваться интуитивно понятными элементами интерфейса, как: кнопка с переключателем состояния (тогл-свитч); выпадающий список параметров; ручной ввод значений.

Некоторые из параметров имеют подсказки, которые отмечены специальным символом ⓘ .Для просмотра подсказки необходимо навести курсор на символ, после чего отобразится всплывающее окно с текстом подсказки, которое будет оставаться видимым до тех пор, пока курсор наведен на выбранный текст.

Для сохранения выбранных параметров необходимо нажать на кнопку «Сохранить» в нижнем углу, кнопка «Сбросить изменения» в нижнем углу позволяет сбросить все введенные значения (Рисунок 4).

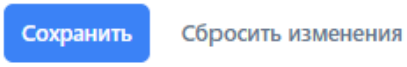

Рисунок 4

Также, в нижней части Системы обозначена область сведений об авторизованном пользователе и кнопка, с помощью которой можно свернуть или развернуть меню раздела.

4.4.Интеграция с ИБ-инфраструктурой:

Таблица 5

Параметр интеграции	Значение / Описание	Реализация	Используемые компоненты
Формат событий	CEF (Common Event Format) для единого парсинга в SIEM/SOAR	Встроенный сериализатор на базе моделей данных	pydantic, json, orjson
Поля события	Timestamp, source_hostname event_id, content_type(prompt/result), Triggered Validator, Risk Score, Action (Block/Pass)	Генерация на лету при срабатывании валидатора	uuid_utils, datetime, pydantic, mmh3
Транспорт	Syslog (UDP/TCP)	Асинхронная отправка без блокировки основного потока	aiohttp, httpx, httpcore, websockets
Частота отправки	Реальное время (по событию/алерту)	Event-driven архитектура + буферизация	asyncio, uvloop, watchdog, tenacity
Совместимость с SIEM/SOC	MaxPatrol, KUMA, Splunk, Elastic Security, отечественные платформы	Стандарт CEF	requests-toolbelt, certifi, urllib3, zstandard

5. Этапы жизненного цикла

Жизненный цикл Программного продукта обеспечивается в соответствии с требованиями ГОСТ Р ИСО/МЭК 12207-2010 «Информационная технология. Системная и программная инженерия. Процессы жизненного цикла программных средств».

В процессе разработки программного продукта принимала участие команда специалистов следующей позиции/квалификации:

- Владелец продукта;
- Системные аналитики (2 человека);
- Технический руководитель команды разработки;
- Команда разработки (2 человека);
- Инженеры тестировщики (1 человек);
- Инженеры по разработке и эксплуатации (DevOPS) (1 человек).

Требования к продукту формируются на основе:

- нормативных актов в области ИБ;
- функциональных запросов заказчиков;
- результатов внутренних ревизий и аудитов.

Все требования регистрируются в системе отслеживания задач, проходят анализ на тестируемость, прослеживаемость и влияние на ИБ среду функционирования.

5.1.Сбор и подготовка данных:

- Инфраструктура подготовки данных.
- Источники данных (логи, промпты, ответы LLM).
- Датасеты для обучения ML-моделей.

5.2.Разработка модели, обучение и тестирование:

- Оценка защищённости и тестирование безопасности ИИ.
- Тестирование модели на стенде разработки.
- Инфраструктура обучения (GPU/CPU).
- Модель в стадии обучения (финализация весов).

Тестирование проводится как вручную, так и автоматизировано с использованием необходимых инструментов проверки и бизнес-логики. Все результаты фиксируются в системе отслеживания задач.

При выявлении дефектов выпускается исправление в виде патча или включается в следующую версию продукта.

5.3.Эксплуатация модели и интеграция:

- Проверка данных и промптов на входе в Продакшн модель.
- Проверка выходных данных на утечки.
- Мониторинг инцидентов в ИИ и подключение SOC.
- Детектирование аномалий в телеметрии модели при инференсе.
- Обработка входных и выходных данных в реальном времени.

5.4.Оценка защищённости и соблюдение требований:

- Непрерывная оценка защищённости модели.
- Тестирование безопасности ИИ на этапе эксплуатации.

5.5.Техническая поддержка:

Техническая поддержка осуществляется в соответствии с практиками управления ИТ-услугами (в том числе на основе принципов ITIL, NIST SP 800-61, «Computer Security Incident Handling Guide») и внутренними регламентами.

Цель поддержки – обеспечение стабильной, безопасной и эффективной эксплуатации продукта в условиях реальных ИБ-инфраструктур заказчиков.

В рамках поддержки предоставляются следующие услуги:

- консультации по установке, настройке и эксплуатации программного продукта;
- диагностика и устранение выявленных неисправностей (багов, сбоев, некорректного поведения);

- выпуск исправлений и обновлений (в том числе срочных патчей при критических уязвимостях);
- актуализация эксплуатационной и справочной документации;
- содействие в интеграции с внешними системами.

Услуги технической поддержки предоставляются только при действующем договоре поддержки в течение указанного календарного периода.

Запросы на техническую поддержку осуществляются по адресу электронной почты:

- ProductOffice@innostage-group.ru

Все письма автоматически регистрируются в системе отслеживания задач. Ответ предоставляется в срок, согласованный в договоре поддержки.

Режим работы службы поддержки:

Базовый уровень поддержки: понедельник–пятница, с 9:00 до 18:00 по Московскому времени.

Реакция на критические инциденты (P1): обеспечивает внережимное сопровождение (включая выходные и праздничные дни) при наличии в договоре условия «24×7 для критических систем».

Сроки первоначального ответа:

- P1 (критический) – не позднее 2 часов с момента регистрации;
- P2 (высокий) – не позднее 8 рабочих часов;
- P3 (средний) – не позднее 1 рабочего дня;
- P4 (низкий / информационный) – не позднее 3 рабочих дней.

Поддержка обеспечивается сертифицированными специалистами, обладающими смежными компетенциями и покрывающими полный цикл сопровождения:

- Обработка запросов от заказчиков;
- Диагностика и первичный анализ инцидентов;
- Формирование отчётов и рекомендаций по устранению проблем;
- Аудит безопасности и расследования инцидентов ИБ;
- Исправления и доработки по результатам запросов;
- Тестирование и верификация патчей;
- Поддержка совместимости продукта с целевыми платформами.

Технические специалисты владеют стеками C# (.NET 6+), JavaScript, PostgreSQL, Redis, Nginx, Python и пр., проходят регулярное обучение по актуальным угрозам ИБ, стандартам защищённой разработки (Secure SDLC) и требованиям нормативных актов (в т.ч. ФСТЭК, ФСБ).

5.6.Совершенствование продукта:

Модернизация Программного продукта осуществляется с учётом:

- изменений в законодательстве РФ в сфере ИБ, ИИ и пр.;
- стратегии развития продукта;
- запросов заказчиков в системе отслеживания задач.

6. Технические требования

6.1.Поддерживаемые операционные системы:

- Astra Linux Special Edition / Common Edition.
- RedOS «Смоленск» / «Орёл» / «Муром».
- Alt Linux «Альт Сервер» / «Альт Рабочая станция».

- Ubuntu 20.04 LTS и совместимые дистрибутивы Linux.
- ОС семейства Redhat, Debian.

6.2. Аппаратные требования (минимальные):

- Процессор: 32 ядра, 2.0 ГГц.
- Оперативная память: 32 ГБ.
- Дисковое пространство: 512 ГБ.
- GPU: NVIDIA с поддержкой CUDA 13+, 12 ГБ Видео Памяти.
- Сетевой интерфейс: 10 Гбит/с.

6.3. Аппаратные требования (рекомендуемые):

- Процессор: 48 ядер, 3.0 ГГц.
- Оперативная память: 64 ГБ.
- Дисковое пространство: 1024 ГБ SSD.
- GPU: NVIDIA с поддержкой CUDA 13+, 24 ГБ Видео Памяти.
- Сетевой интерфейс: 10 Гбит/с.

6.4. Программные зависимости:

- docker – версия не ниже 20.10.12.
- docker-compose – версия не ниже 2.4.1.
- bash.

7. Эксплуатационные характеристики

7.1. Производительность:

- Обработка запросов: до 50 запросов в секунду.
- Задержка обработки (latency): не более 6000 мс.
- Доступность: 99.9% (при конфигурации с резервированием).

7.2. Масштабируемость:

- Горизонтальное масштабирование через балансировщик нагрузки.
- Контейнеризация: Docker.

7.3. Резервное копирование и восстановление:

- Возможность восстановления из резервной копии через CLI.
- Интервал резервного копирования: настраивается администратором.

8. Требования к безопасности

8.1. Защита от несанкционированного доступа:

- Аутентификация администраторов через локальную базу.
- Журналирование всех действий администраторов.

8.2.Защита данных:

- Шифрование конфиденциальных данных в хранилище.
- Защита каналов связи (TLS 1.2+).

8.3.Аудит и мониторинг:

- Полное логирование событий безопасности.
- Интеграция с SIEM-системами.

9. Условия эксплуатации

9.1.Квалификация персонала:

- Базовые навыки администрирования ОС семейства Linux.
- Базовые навыки работы с сетевой инфраструктурой.
- Базовые навыки работы с ML/LLM (для расширенной настройки).
- Базовые навыки работы с средствами мониторинга ИБ.

10.Контактная информация

- Наименование организации: ООО «Инностейдж ЦР», генеральный директор Гузаиров Айдар Фаилевич.
- Фактический адрес размещения разработчиков, службы поддержки и инфраструктуры разработки: 420015, Республика Татарстан, г. Казань, ул. Подлужная, д. 60.
- Адрес сайта: https://inno-dev.ru/ai-dr