Innostage
центр
разработок

# РУКОВОДСТВО ПО УСТАНОВКЕ И НАСТРОЙКЕ

ПРОГРАММНЫЙ ПРОДУКТ
**«Защита ИИ. AIDR»**
версия 1.0

2026 г.

# Оглавление

# 1. Введение

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 предназначен для специалистов, обеспечивающих установку и настройку Программного продукта, и содержит системные требования к развертыванию продукта, а также инструкции по установке.

# 2. Системные требования

## 2.1.Системные требования к оборудованию:

Системные требования к окружению представлены в таблице ниже в Таблице 1.

Таблица 1

| ОПЕРАЦИОННАЯ СИСТЕМА | АРХИТЕКТУРА | ЯДРО LINUX | ТРЕБОВАНИЕ К ПРИКЛАДНОМУ ПО | МИНИМАЛЬНЫЕ ТРЕБОВАНИЯ КОНФИГУРАЦИИ | КОЛИЧЕСТВО СЕРВЕРОВ |
|---|---|---|---|---|---|
| **linux / amd64** | x86_64 | >= 4.0 | Wget, curl, tar, docker, docker compose | 16 ЯДРА / 32 ГБ оперативной памяти / 500 ГБ жесткого диска | 1 |

**Примечание:** требуется ядро ОС семейства Redhat, Debian, Ubuntu версии 4.0 и выше.

## 2.2.Требования к персоналу:

Для установки и администрирования системы персонал должен обладать следующими навыками и компетенциями:

- Управление и настройка Web-серверов.
- Управление и настройка системам контейнеризации.
- Управление и настройка системам на базе ядра Linux.
- Понимание принципов работы больших языковых моделей и систем их запуска.

## 2.3.Общие требования:

1. Доступ к модели GUARD_LLM, запущенной локально либо на стороннем сервисе. (endpoint url провайдера LLM и токен доступа(опционально)).
2. Доступ к защищаемой модели или LLM-шлюзу.
3. Настроен сервер для разворота системы:

- docker - версия не ниже 20.10.12.
- docker-compose - версия не ниже 2.4.1.
- bash.
- nano или vim.

4. Открыт доступ в Internet (опционально).
5. Свободные порты, необходимые системе.
6. Обеспечить наличие свободного места в каталоге, где будут размещаться все файлы приложения – ‘/opt’, далее в тексте упоминаемая как рабочая директория *$WORK-DIR*.

7. Существует учетная запись пользователя, от имени которого будут выполняться все действия по развороту и обслуживанию приложения:

*Пользователь должен иметь полные права на каталог $WORK-DIR и права для запуска docker и docker-compose (как это сделать, описано в инструкции по ссылке: https://docs.docker.com/engine/install/linux-postinstall/ (раздел Manage Docker as a non-root user)), либо права супер-пользователя (sudo).*

8. Дистрибутив должен быть скачан с официальных ресурсов ООО «Инностейдж ЦР». Дистрибутив содержит архив и файлы, необходимые для запуска приложения.

# 3. Установка программного продукта

## 3.1. Установка:

Установка осуществляется посредством инсталлятора «setup-aidr-{версия продукта}.run», который в интерактивном режиме запрашивает параметры установки у пользователя.

Инсталлятор поставляется в незашифрованном и зашифрованном форматах, если получен зашифрованный формат, то ключ запрашивается у специалистов ООО «Инностейдж ЦР» и передается отдельным каналом связи.

На основе ответов инсталлятор формирует и загружает docker-образы, docker-compose файлы сервисов и выполняет разворачивание с помощью инструментария docker compose.

Для установки необходимо выполнить следующие действия на сервере (примеры приведены для deb ОС):

1. Установить вспомогательные пакеты:

```
apt install docker.io
apt install docker-compose-v2
```

2. Перейти в каталог с файлом «setup-aidr-*.run», выдайте права на исполнение пакета «setup-aidr -*.run»:

```
chmod +x setup-aidr -*.run
```

3. Выполните запуск инсталлятора:

```
sudo -i ./setup-aidr -*.run
```

4. Инсталлятором будут заданы вопросы по конфигурации сервисов. Необходимо указать запрашиваемые параметры (см. справочник по переменным окружения из «Руководства администратора»).

5. После завершения скрипта будет создан файл «.env» в котором можно в дальнейшем изменять заданные значения.

6. После завершения скрипта установка продукта «ЗАЩИТА ИИ. AIDR» будет завершена. Система будет доступна по следующим портам:

   a. 8008 – порт для подключения приложений к сервису прокси «ЗАЩИТА ИИ. AIDR»

   b. 80 – порт для UI административной консоли (по умолчанию доступ по admin/admin)

Полный процесс установки представлен на Рисунок 1:

```
Verifying archive integrity...  100%   SHA256 checksums are OK. All good.
Decrypting and uncompressing AIDR install script...
     0% enter AES-256-CBC decryption password:
*** WARNING : deprecated key derivation used.
Using -iter or -pbkdf2 would be better.

::: Installing and configuring AIDR

 -> To start using the product, first accept the license agreement

TODO add license
Enter "y" to accept, "n" to cancel: y
[ 1/4 ] Checking the environment
 -> Successfully checked

[ 2/4 ] Copying the installer distribution
 -> Successfully copied

[ 3/4 ] Generate config

 -> Please provide configuration parameters:
 -> Configuring Guardrails AI Service:
 -> Configuring Gateway Service:
Guard llm base URL [default: https://openrouter.ai/api/v1]:
Guard llm api key [default: ]: test
URL of the protected LLM [default: http://localhost:8080/v1]:
Api key of the protected LLM [default: ]: test
 -> Configuring Admin UI Service:
Domain name or IP address of the administration server [default: 10.70.54.161]:
 -> Successfully generated

[ 4/4 ] Loading the docker images to cache
Loaded image: registry.innostage-group.ru/deprod-docker-aidr/nemoguardrails_service:1.0.0
Loaded image: registry.innostage-group.ru/deprod-docker-aidr/admin_ui_service:1.0.0
Loaded image: registry.innostage-group.ru/deprod-docker-aidr/ml_grai_service:1.0.0
Loaded image: registry.innostage-group.ru/deprod-docker-aidr/static_ag_service:1.0.0
Loaded image: registry.innostage-group.ru/deprod-docker-aidr/static_lg_service:1.0.0
 -> Successfully loaded

::: Installation completed!
>>> To change the default configuration values:
>>> edit the /opt/test/aidr/.env file

>>> To start the application:
>>> go to the directory /opt/test/aidr
>>> docker compose up -d
```

*Рисунок 1 – Установка продукта*

### 3.2.Чек-лист для установки «ЗАЩИТА ИИ. AIDR»:

- ☐ Проверить URL сервиса «ЗАЩИТА ИИ. AIDR»
- ☐ Проверить URL защищаемой и защитной LLM (Guard LLM, Protected Main LLM)
- ☐ Заполнить API-ключи LLM (Guard LLM, Protected Main LLM), если используются

## 4. Эксплуатация и обслуживание программного продукта

### 4.1.Эксплуатация и обслуживание с помощью веб-консоли:

Подключение к веб-консоли производится по url: [ip-адресс/имя сервера AIDR]:80.
По умолчанию доступ открыт под учетной записью admin (пароль по умолчанию: admin).

В веб консоли доступна настройка политики валидаторов:

1) Настройки валидатора «Персональные данные» представлены на Рисунок 2

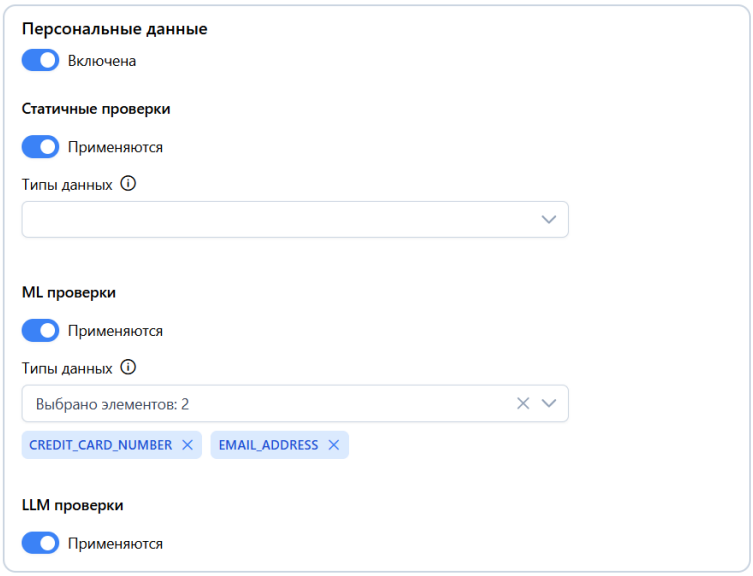


*Рисунок 2 - Настройки валидатора «Персональные данные»*

2) Настройки валидатора «Секреты» представлены на Рисунок 3

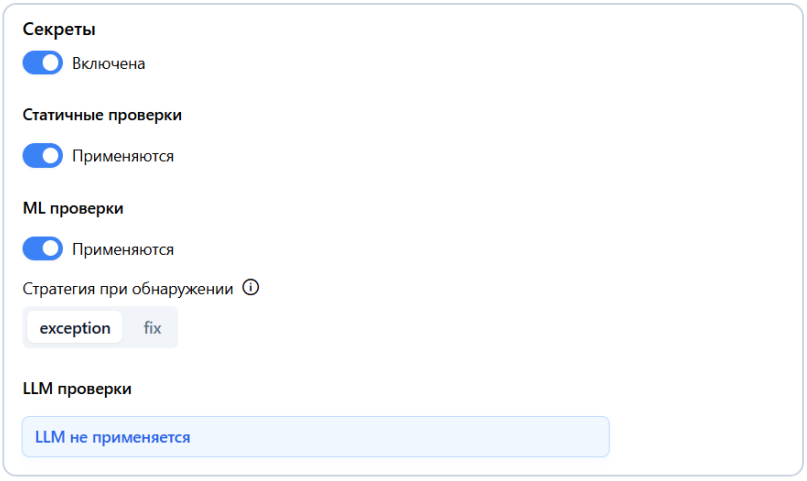


*Рисунок 3 - Настройки валидатора «Секреты»*

3) Настройки валидатора «Контроль темы» представлены на Рисунок 4

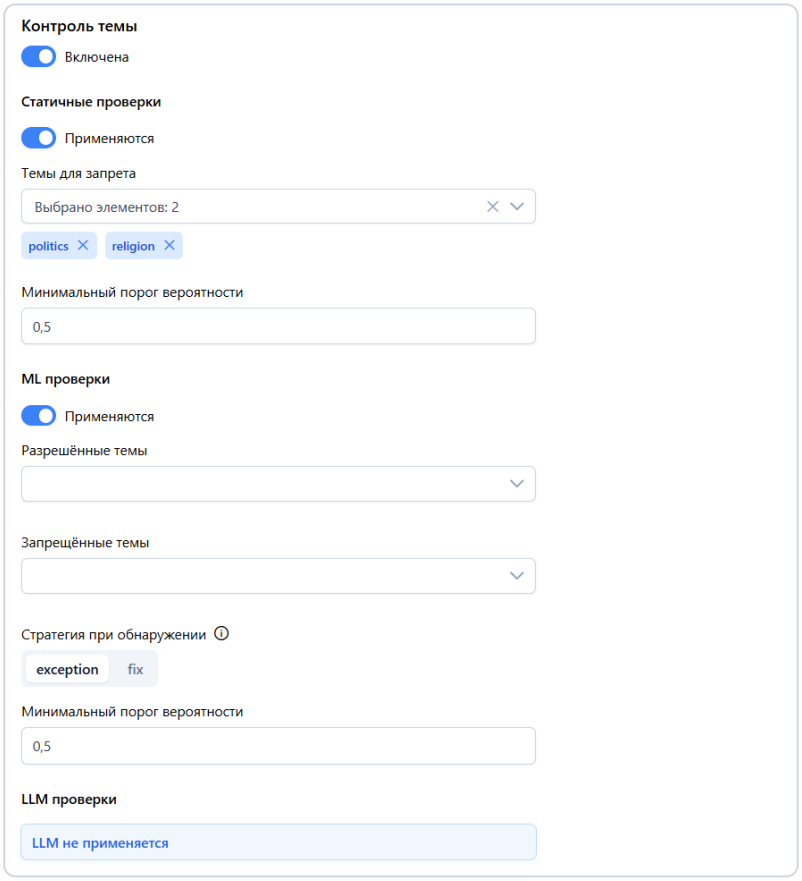


*Рисунок 4 - Настройки валидатора «Контроль темы»*

4) Настройки валидатора «Токсичный контент» представлены на Рисунок 5

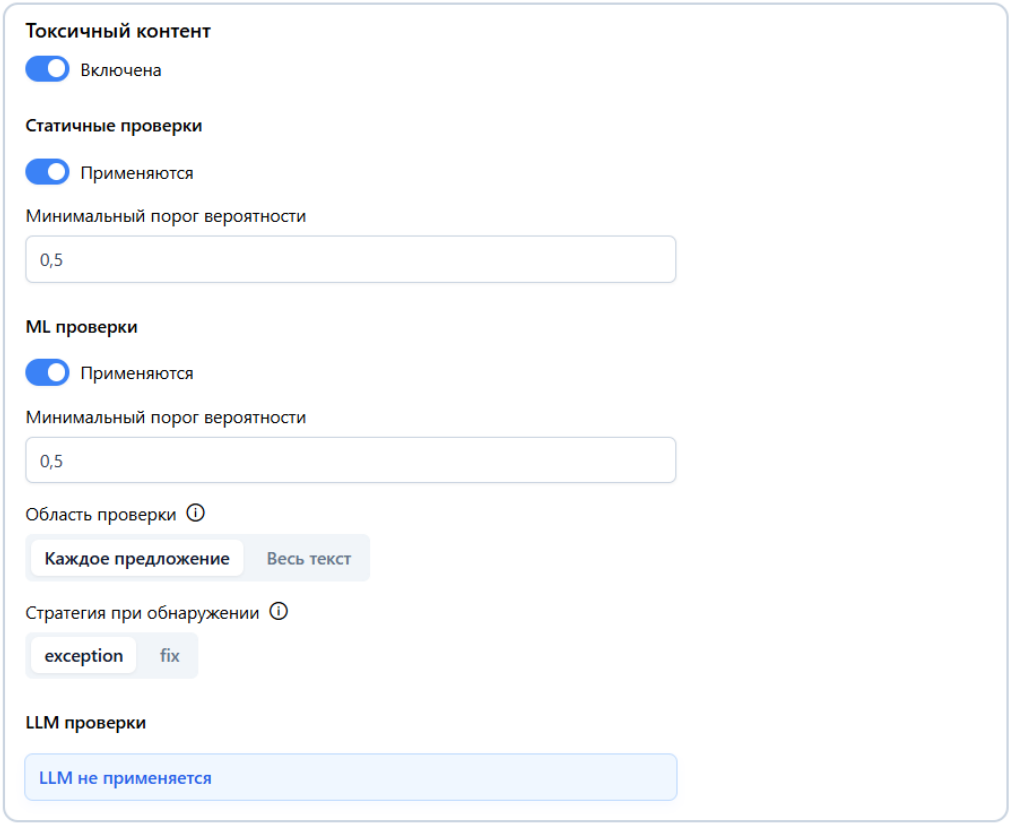


*Рисунок 5 - Настройки валидатора «Токсичный контент»*

5) Настройки валидатора «Небезопасный контент» представлены на Рисунок 6

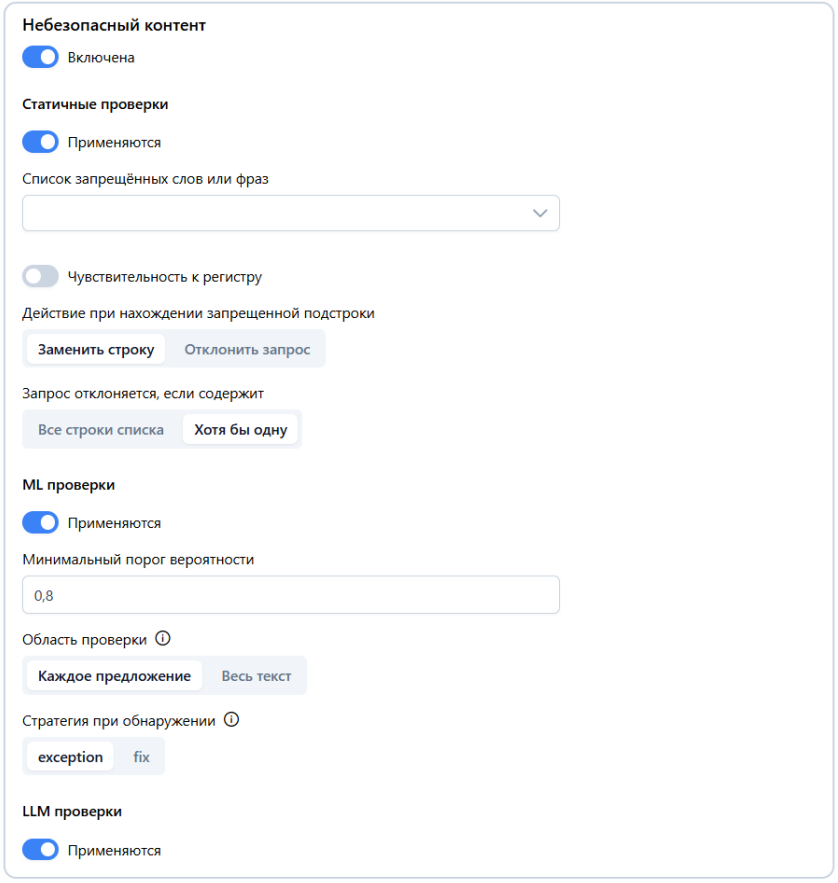


*Рисунок 6 - Настройки валидатора «Небезопасный контент»*

6) Настройки валидатора «Попытка jailbreak/prompt injection» представлены на Рисунок 7

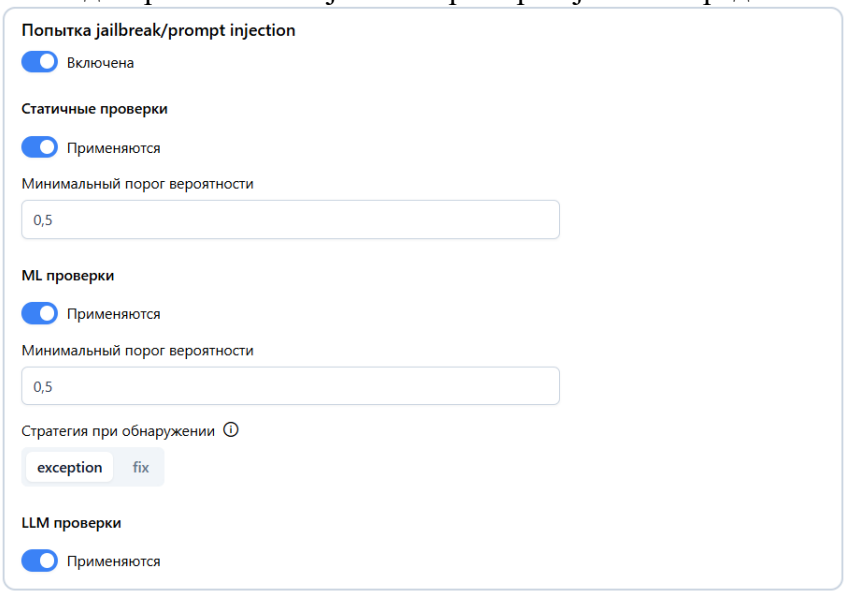


*Рисунок 7 - Настройки валидатора «Попытка jailbreak/prompt injection»*

7) После внесения изменений в политику можно их сохранить или сбросить нажав на соответствующую кнопку Рисунок 8


*Рисунок 8 – Сохранение/сброс настроек политики*

### 4.2.Эксплуатация и обслуживание с помощью командной строки:

Управление модулями происходит посредством команд docker и docker-compose.
Для остановки сервисов выполните команду в корневом каталоге модуля:

```
Docker compose stop
```

Для запуска сервисов:

```
Docker compose up -d
```

Для просмотра логов:

```
Docker compose logs <имя сервиса>
```

Управление конфигурациями модулей производится через смену значений в переменных окружения. Полный перечень доступных к конфигурированию переменных окружения описан в п.5.Справочник переменных окружения.

## 5. Справочник переменных окружения

Полный справочник по всем переменным окружения, используемым в проекте.

### 5.1.Общие принципы:

Таблица 2

| ПРИОРИТЕТ | ИСТОЧНИК |
|---|---|
| 1 | Переменные окружения (export VAR=value) |
| 2 | Файл .env сервиса |
| 3 | Значения по умолчанию в коде |

### 5.2.Корневой .env:

Файл расположен в корне проекта (./.env). Используется всеми сервисами через env_file в docker-compose.yml.

### 5.3.Общие настройки:

Таблица 3

| ПЕРЕМЕННАЯ | ОПИСАНИЕ | ПО УМОЛЧАНИЮ | ОБЯЗАТЕЛЬНАЯ |
|---|---|---|---|
| PYTHONHTTPSVERIFY | Отключение проверки SSL (0 = отключено) | 0 | Нет |
| DISABLE_PII_SERVICE | Отключение PII сервиса | false | Нет |
| DISABLE_JAILBREAK_SERVICE | Отключение Jailbreak сервиса | false | Нет |

### 5.4.API-ключи:

Таблица 4

| ПЕРЕМЕННАЯ | ОПИСАНИЕ | ОБЯЗАТЕЛЬНАЯ |
|---|---|---|
| GUARD_LLM_API_KEY | Guard LLM API ключ для Guard LLM | Нет |
| GUARD_LLM_URL | URL Guard LLM | https://LOCAL_GUARD_LLM_URL/api/v1 |
| MAIN_LLM_API_KEY | API ключ основной LLM | Нет(указать если требуется) |

| ПЕРЕМЕННАЯ | ОПИСАНИЕ | ОБЯЗАТЕЛЬНАЯ |
|---|---|---|
| MAIN_MODEL_BASE_URL | URL основной LLM | https:// LOCAL_PROTECTED_LLM_URL/v1 |

### 5.5.AI Detection and Response Proxy:

Таблица 5

| ПЕРЕМЕННАЯ | ОПИСАНИЕ | ПО УМОЛЧАНИЮ | ОБЯЗАТЕЛЬНАЯ |
|---|---|---|---|
| DEFAULT_CONFIG_ID | ID конфигурации Rails | aidr | Нет |
| HOST | Хост для binding | 0.0.0.0 | Нет |
| PORT | Порт NemoGuardrails | 8008 | Нет |

### 5.6.SIEM / Syslog:

Таблица 6

| ПЕРЕМЕННАЯ | ОПИСАНИЕ | ПО УМОЛЧАНИЮ |
|---|---|---|
| SIEM_SYSLOG_ENABLE | Включение Syslog | false |
| SIEM_SYSLOG_ADDR | Адрес Syslog сервера | SYSLOG_SERVER_IP:PORT |
| SIEM_SYSLOG_FACILITY | Syslog facility | local0 |
| SIEM_SYSLOG_SOCKTYPE | Тип сокета (tcp/udp) | tcp |
| SIEM_DEVICE_PRODUCT | Название продукта в SIEM | AIDR |
| SIEM_DEVICE_VERSION | Версия продукта в SIEM | v1.0 |

### 5.7.Сетевые настройки ml_grai_service:

Таблица 7

| ПЕРЕМЕННАЯ | ОПИСАНИЕ | ПО УМОЛЧАНИЮ | DOCKER | VENV |
|---|---|---|---|---|
| PORT | Порт сервера | 8006 | Да | Да |
| HOST | Хост для binding | 0.0.0.0 | Да | Да |

### 5.8.Пороги валидаторов:

Таблица 8

| ПЕРЕМЕННАЯ | ОПИСАНИЕ | ПО УМОЛЧАНИЮ |
|---|---|---|
| JAILBREAK_DEFAULT_THRESHOLD | Порог Jailbreak (0.0–1.0) | 0.5 |
| TOPIC_DEFAULT_THRESHOLD | Порог Topic (0.0–1.0) | 0.3 |
| TOXIC_DEFAULT_THRESHOLD | Порог Toxic (0.0–1.0) | 0.5 |
| NSFW_DEFAULT_THRESHOLD | Порог NSFW (0.0–1.0) | 0.7 |

### 5.9.Контроль Персональных данных(PII) настройки:

Таблица 9

| ПЕРЕМЕННАЯ | ОПИСАНИЕ | ПО УМОЛЧАНИЮ |
|---|---|---|
| PII_SKIP_ENTITY_TYPES | Типы сущностей для пропуска | LOCATION, GPE, URL,DATE_TIME,IP_ADDRESS |
| PII_SENSITIVE_ENTITY_TYPES | Чувствительные типы сущностей | EMAIL_ADDRESS, PHONE_NUMBER,PERSON, CREDIT_CARD, SSN, PASSWORD, API_KEY |

### 5.10. Контроль секретов (Secrets) настройки:

Таблица 10

| ПЕРЕМЕННАЯ | ОПИСАНИЕ | ПО УМОЛЧАНИЮ |
|---|---|---|
| SECRETS_PATTERNS | Кастомные regex паттерны (формат: regex:name) | [] |

### 5.11. Логирование:

Таблица 11

| ПЕРЕМЕННАЯ | ОПИСАНИЕ | ПО УМОЛЧАНИЮ |
|---|---|---|
| LOG_LEVEL | Уровень логирования | INFO |

### 5.12. static_lg_service:

Файл: services/static_lg_service/config.py

Таблица 12

| ПЕРЕМЕННАЯ | ОПИСАНИЕ | ПО УМОЛЧАНИЮ | ОБЯЗАТЕЛЬНАЯ |
|---|---|---|---|
| PORT | Порт сервера | 8004 | Нет |
| HOST | Хост для binding | 0.0.0.0 | Нет |
| LOG_LEVEL | Уровень логирования | INFO | Нет |
| RELOAD_MODE | Режим разработки (hot reload) | false | Нет |
| SCANNER_TIMEOUT | Таймаут сканеров (сек) | 30 | Нет |
| TOKEN_LIMIT | Лимит токенов | 4096 | Нет |
| DEFAULT_INPUT_SCANNERS | Сканеры по умолчанию | secrets,token_limit | Нет |
| APP_VERSION | Версия приложения | 1.0.0 | Нет |

### 5.13. static_ag_kt_service:

Использует переменные из корневого .env и внутренние настройки Kotlin-приложения.

Таблица 13

| ПЕРЕМЕННАЯ | ОПИСАНИЕ | ПО УМОЛЧАНИЮ |
|---|---|---|
| JAVA_OPTS | JVM опции | -Xms512m -Xmx1536m -XX:+UseG1GC |

### 5.14. AI Detection and Response Proxy:

Использует переменные из корневого .env и дополнительные через environment: в docker-compose.yml.

Таблица 14

| ПЕРЕМЕННАЯ | ОПИСАНИЕ | ПО УМОЛЧАНИЮ |
|---|---|---|
| RAILS_CONFIG_PATH | Путь к конфигурации Rails | /app/config/aidr |
| DEFAULT_CONFIG_ID | ID конфигурации | aidr |
| HOST | Хост для binding | 0.0.0.0 |
| PORT | Порт сервера | 8008 |
| OMP_NUM_THREADS | Число потоков OpenMP | 2 |
| REQUESTS_TIMEOUT | Таймаут HTTP-запросов (сек) | 5 |
| GUARDRAILS_AI_TIMEOUT | Таймаут к Guardrails AI (сек) | 30 |
| LLM_GUARD_TIMEOUT | Таймаут к LLM-Guard (сек) | 5 |
| ANGRYSCAN_TIMEOUT | Таймаут к AngryScan (сек) | 3 |
| NEMO_FAST_MODE | Быстрый режим NemoGuardrails | false |
| NEMO_ENABLE_CACHING | Включение кэша | true |

| ПЕРЕМЕННАЯ | ОПИСАНИЕ | ПО УМОЛЧАНИЮ |
|---|---|---|
| PYTHONHTTPSVERIFY | Отключение SSL проверки | 0 |

## 6. Приложение 1: Перечень используемых сокращений

| СОКРАЩЕНИЕ | ПОЛНОЕ НАИМЕНОВАНИЕ |
|---|---|
| **БД** | База данных |
| **ИБ** | Информационная безопасность |
| **ИТ** | Информационные технологии |
| **НМЖД** | Накопитель на жестких магнитных дисках |
| **ОЗУ** | Оперативное запоминающее устройство |
| **ОС** | Операционная система |
| **ПО** | Программное обеспечение |
| **Продукт** | Программный продукт «Защита ИИ. AIDR» |
| **СУБД** | Система управления базой данных |
| **ЦПУ** | Центральное процессорное устройство |
| **API** | Application Programming Interface |
| **IP** | Internet Protocol |
| **MS** | Microsoft |
| **SIEM** | Security Information and Event Management |
| **SQL** | Structured Query Language |
| **URL** | Uniform Resource Locator |
| **XML** | Extensible Markup Language |

## 7. Приложение 2: Перечень терминов и определений

| ТЕРМИН | ОПРЕДЕЛЕНИЕ |
|---|---|
| **Инцидент ИБ** | Факт нарушения и (или) прекращения функционирования информационного ресурса и (или) нарушения безопасности, обрабатываемой таким информационным ресурсом информации, в том числе произошедший в результате компьютерной атаки |
| **Объект структуры организации** | Юридические адреса, структурные подразделения, адреса, подразделения, рабочие группы, работники |

# 8. Приложение 3: Журнал регистрации изменений

| **Журнал регистрации изменений** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Изм. | Номера листов (страниц) | | | | Всего листов (страниц) в док. | Номер док. | Подп. | Дата |
| | изме-ненных | заменен-ных | новых | аннулиро-ванных | | | | |
| | | | | | | | | |