Годовикова 9с17
7 495 234 2222
info@skala-

Описание эксплуатации программного обеспечения для электронно-вычислительных машин «Скала^р Спектр ИИ»

Москва 2026 г.

Годовикова 9с17
7 495 234 2222
info@skala-

Содержание

Годовикова 9с17
7 495 234 2222
info@skala-

Термины и определения

Термин	Значение
Agent	Агент - Автономная AI-система, способная выполнять последовательность действий для достижения цели, используя инструменты (tools)
AI	Artificial Intelligence – искусственный интеллект
AIOps	Artificial Intelligence for IT Operations — это использование искусственного интеллекта, машинного обучения (ML) и больших данных для автоматизации и оптимизации ИТ-операций
API	Application Programming Interface - Интерфейс программирования приложений
Chatbot	AI-ассистент для диалога с пользователем
Chain of Thought	Цепочка рассуждений, метод пошаговой генерации ответа LLM
Citations	Ссылки на источники в ответах RAG-системы
Context Precision	Метрика Ragas, измеряющая релевантность извлечённого контекста
Context Recall	Context Recall - Метрика Ragas, измеряющая полноту извлечённого контекста |
DevOps	Development & operations — методология автоматизации технологических процессов сборки, настройки и развёртывания программного обеспечения
IP-адрес	Уникальный сетевой адрес узла в компьютерной сети, построенной на основе стека протоколов TCP/IP.
Embedding	Embedding - Векторное представление текста или изображения|
Faithfulness	Метрика Ragas, измеряющая фактическую корректность ответа
Inference	Процесс генерации ответа LLM
JWT	JSON Web Token - Стандарт токенов для аутентификации

Годовикова 9с17
7 495 234 2222
info@skala-

LDAP	Lightweight Directory Access Protocol – «легкорасширяемый протокол доступа к каталогам» – протокол прикладного уровня для доступа к службе каталогов X.500. Относительно простой протокол, использующий TCP/IP и позволяющий производить операции аутентификации (bind), поиска (search) и сравнения (compare), а также операции добавления, изменения или удаления записей.
LLM	Large Language Model - большая языковая модель. Тип нейросетей, обученных на огромных объемах текста для понимания, обработки и генерации человекоподобного языка
Low-code	Подход к разработке приложений с минимальным кодированием, через визуальный интерфейс
ML	Машинное обучение (Machine Learning, ML) — это область искусственного интеллекта, разрабатывающая алгоритмы, способные обучаться на данных, находить закономерности и делать прогнозы без явного программирования под каждую задачу
MLOps	Набор практик, нацеленных на надежное и эффективное развертывание и поддержание моделей машинного обучения на производстве. Слово является смесью слов "машинное обучение" (ML) и практик непрерывной разработки — DevOps в области программного обеспечения
MTBF	Mean Time Between Failurest - Среднее время между отказами
NTP	Сетевой протокол для синхронизации внутренних часов компьютера с использованием сетей с переменной латентностью.
OAuth 2.0	Протокол авторизации
Observability	Наблюдаемость системы (журналы, метрики, traces)
OCR	Optical Character Recognition — оптическое распознавание символов. Это технология преобразования изображений (сканов, фото, PDF) в редактируемый машиночитаемый текстовый формат
OIDC	OpenID Connect - Протокол аутентификации поверх OAuth 2.0

PostgreSQL	Свободная объектно-реляционная система управления базами данных. Существует в реализациях для множества UNIX-подобных платформ.
Prompt	Текстовый запрос к LLM |
Prompt Engineering	Процесс оптимизации промптов для улучшения качества ответов LLM
RAG	Retrieval-Augmented Generation — генерация, дополненная поиском — это технология AI, улучшающая ответы LLM путем поиска информации во внешних источниках (документы, базы данных) перед генерацией ответа
RBAC	Role-Based Access Control - Управление доступом на основе ролей |
RPO	Recovery Point Objective - Целевая точка восстановления (максимально допустимая потеря данных)
RTO	Recovery Time Objective - Целевое время восстановления (максимально допустимое время простоя)
SLA	Service Level Agreement - Соглашение об уровне обслуживания
Span	Единица трассировки (часть trace)
SSL	Secure Sockets Layer – уровень защищенных сокетов – криптографический протокол, который подразумевает более безопасную связь. Он использует асимметричную криптографию для аутентификации ключей обмена, симметричное шифрование для сохранения конфиденциальности, коды аутентификации сообщений для целостности сообщений.
SSO	Single Sign-On - Единая точка входа
Tool Calling	Возможность LLM вызывать внешние инструменты (API, функции)
TPOT	Time Per Output Token - Время генерации одного токена
Trace	Полное отслеживание запроса (цепочка spans)
TTFT	Time To First Token- Время до первого токена (начала ответа)
Vector Database |	Векторная база данных для хранения embeddings
VLM	Vision Language Model- Мультимодальная модель
Workflow	Пайплайн обработки в Low-code редакторе

Инстанс	Экземпляр, запущенный на одном из серверов сегментов, хранящий и обрабатывающий свою часть данных. Экземпляр класса в объектно-ориентированном программировании.

Сокращения и обозначения

Сокращение	**Наименование**
A2A	Agent-to-Agent (A2A) — открытый стандарт коммуникационного протокола для обеспечения взаимодействия и совместимости между независимыми, потенциально закрытыми AI-агентами
CPU	Central processing unit, центральный процессор.
GPU	Графический процессор (graphics processing unit) — это специализированное электронное устройство, оптимизированное для параллельной обработки данных и выполнения сложных математических вычислений с высокой скоростью
MCP	Model Context Protocol — открытый протокол прикладного уровня для взаимодействия языковых моделей (LLM) с внешними источниками данных и инструментами.
RAM	Random Access Memory, оперативная память
ОС	Операционная система
ПО	Программное обеспечение
ПАК	Программно-аппаратный комплекс, совокупность аппаратных компонент (сервера, коммутаторы доступа и т. п.) и программного обеспечения, решающих некие прикладные задачи (машина баз данных, система управления виртуализацией и так далее)
СУБД	Система управления базами данных
ТЗ	Техническое Задание. Настоящий документ.
ЭВМ	Электронно-вычислительная машина.

Годовикова 9с17
7 495 234 2222
info@skala-

Назначение Системы

Основным предназначением ПО «Спектр ИИ» является централизованное управления ресурсами ПАК (CPU/RAM/GPU) в сценариях использования ИИ-моделей, приложений с ИИ-функциональностью, агентских систем, а также рабочих столов.

ПО «Спектр ИИ» позволит повысить эффективность использования GPU-карт в рамках ПАК, соотнеся вид нагрузки и часть ресурса карты. Также Спектр ИИ сократит время подготовки ПАК для начала работы с моделями, приложениями, агентами и рабочими столами, а также устраняет необходимость в установке и конфигурации дополнительного специализированного ПО поверх ОС.

Интерфейсы взаимодействия

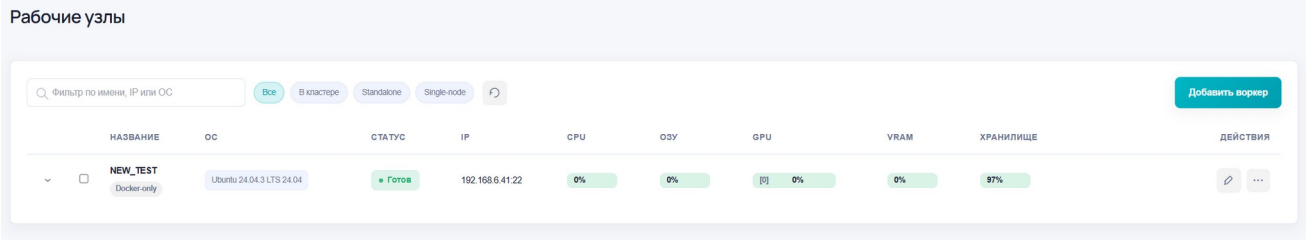
В Системе реализован пользовательский интерфейс позволяющий осуществлять действие в Системе, а также настройку Системы и управление пользователями. Интерфейс имеет русский язык.

В Интерфейсе реализованы следующие разделы:

- Создание контейнеров
- Каталог
- Модели
- Рабочие узлы
- Холст
- Chat UI
- OCR
- Транскрибиация

Среда для конфигурации хостов

Хост — это сервер или виртуальная машина с установленным Docker, на котором развертываются контейнеры. Управление хостами — первый шаг при использовании системы.



Просмотр списка хостов

- Название хоста
- Операционная система
- Статус (Готов / Недоступен)
- IP-адрес / DNS
- CPU (количество ядер, частота)
- RAM (объем)
- GPU модель
- VRAM (объем видеопамяти)
- Дисковое пространство (всего / занято / свободно)

Создание (подключение) хоста

Предварительные условия: Хост должен быть физически доступен по сети, на нем должен быть установлен Docker (или будет установлен автоматически).

Шаги:

1. Нажмите кнопку «Создать хост» или «Подключить хост».
2. Заполните параметры:

- Название — уникальное имя хоста (проверяется системой)
- IP / DNS — сетевой адрес хоста
- SSH порт — порт для SSH-подключения (по умолчанию 22)
- Логин — учетная запись для доступа
- Пароль — пароль учетной записи

3. Нажмите «Подключить».

Что происходит:

- Система выполняет запрос на подключение к хосту
- Создается запись о хосте (даже при ошибках доступа)
- Проверяется конфигурация: SSH-доступ, ОС, пакетный менеджер, Docker
- Выполняется поиск видеокарт
- При необходимости Docker устанавливается автоматически

Результат:

- При успешном подключении — хост появляется в списке со статусом «Готов»
- При ошибке — отображается сообщение с деталями, хост сохраняется в списке с пометкой о недоступности

Информация, отображаемая после подключения:

- Название и версия ОС
- Пакетный менеджер
- Версия Docker

- Список видеокарт (или сообщение «GPU не найдены»)
- CPU, RAM, VRAM, дисковое пространство
- Поддержка технологии MIG (если применимо)

Изменение хоста

Ограничение: Изменению подлежит только название хоста. Остальные параметры определяются при подключении и не редактируются.

Шаги:

1. В списке хостов выберите нужный хост.
2. Нажмите «Изменить».
3. Введите новое уникальное название.
4. Сохраните изменения.

При ошибке: Если имя не уникально, система выдаст сообщение, имя останется прежним.

Удаление (отключение) хоста

Шаги:

1. В списке хостов выберите нужный хост.
2. Нажмите «Удалить».

Варианты:

- Если контейнеров нет — хост удаляется из списка.
- Если контейнеры есть — система запрашивает подтверждение: «На хосте есть запущенные контейнеры. Удалить их вместе с хостом?»

- При согласии: все контейнеры и образы на хосте удаляются, хост исчезает из списка.
- При отказе: удаление отменяется.

Особый случай: Если хост недоступен, контейнеры не могут быть остановлены штатно. Система помечает объекты как недоступные и удаляет их из видимости.

Запрос конфигурации (обновление ресурсов)

Шаги:

1. В списке хостов выберите один или несколько хостов.
2. Нажмите «Проверить конфигурацию».
3. Система выполняет запрос и обновляет изменившиеся показатели.
4. Отображается информация о том, что изменилось.

При ошибке: Если хост недоступен, действие не выполняется, выводится сообщение.

Среда для каталогизации объектов (Каталог)

Каталог — центральное место управления образами и контейнерами. Здесь MLOps-специалист выбирает, загружает и запускает модели и приложения.

Структура каталога

Доступные типы объектов:

- LLM модели — языковые модели для чата и агентов
- Приложения — AI-приложения (RAG, OCR, транскрибация, агентские платформы)
- Рабочие столы — виртуализированные окружения
- Системные контейнеры — ADA Framework, Чат, Холст

- Примеры решений — демонстрационные сервисы

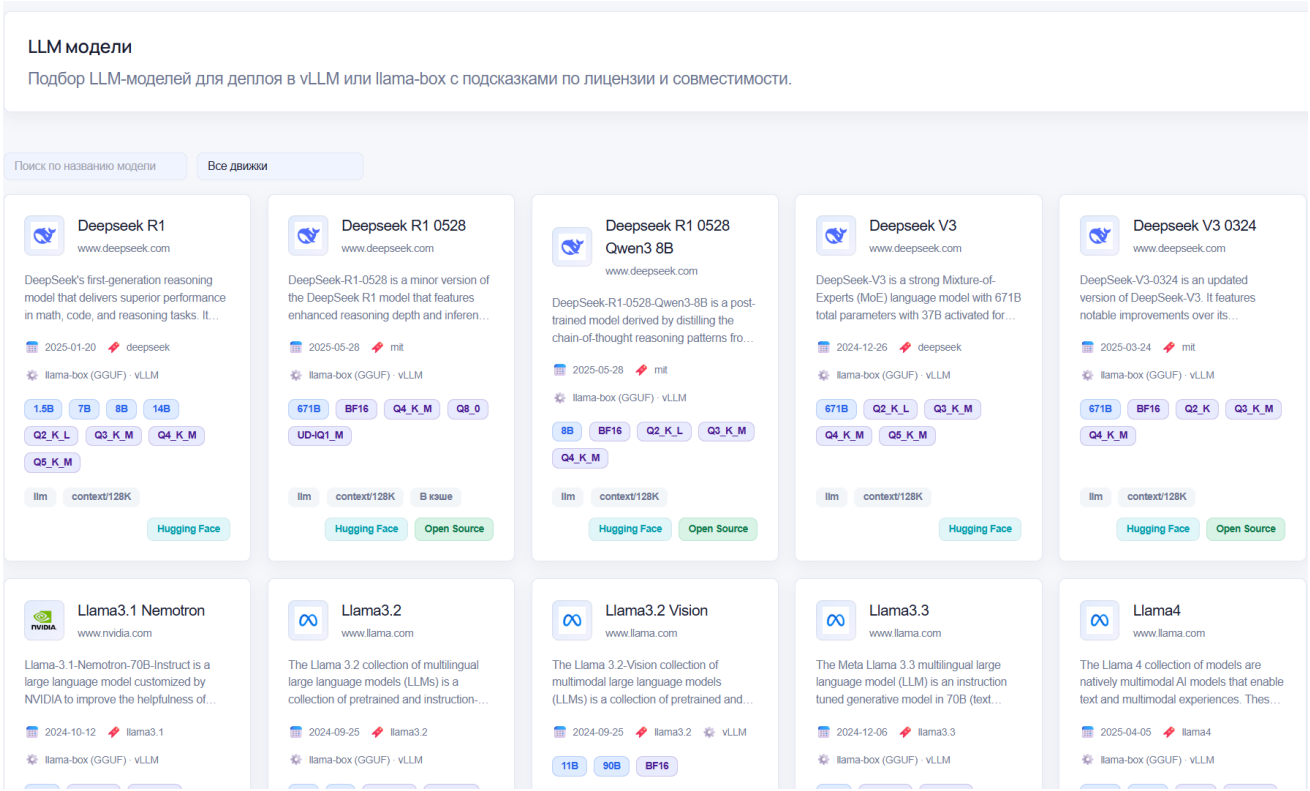
Состояния объектов

Состояние	Описание	Доступные действия
Образ	Метаинформация об объекте, ссылка для скачивания	Удалить, Запустить (создать контейнер)
Контейнер (запущен)	Работающий экземпляр образа	Остановить, Перезапустить
Контейнер (остановлен)	Неактивный экземпляр	Удалить, Запустить
Удален	Полная очистка	Не отображается, восстановление невозможно

Видимость по ролям:

- Администратор, MLOps: видят все объекты во всех состояниях
- Пользователь: видят только запущенные контейнеры, назначенные им

Просмотр моделей



Отображаемая информация для LLM:

- Место хранения (репозиторий)
- Лицензия
- Семейство моделей
- Размерность (количество параметров)
- Temperature, Top P

- Максимальный размер ответа (токены)
- Размер контекстного окна
- Quantization (сжатие)
- max_num_seqs (параллельность запросов)
- Источник загрузки (Hugging Face, корпоративный репозиторий, предустановленный образ)
- Движок инференса (vLLM, SGLang, Ollama, TensorRT-LLM)

Просмотр приложений

Отображаемая информация:

- Дата создания / обновления
- Автор / публикатор
- Размер (ГБ)
- Статус контейнера

Создание (импорт) образа

Предварительные условия: Подключены хосты, настроены репозитории в разделе «Настройки».

Шаги:

1. В Каталоге нажмите «Импортировать образ».
2. Выберите тип объекта (модель / приложение / рабочий стол).
3. Укажите источник:

- Ссылка на Docker registry (с учетными данными при необходимости)
- Локальный файл образа

4. Подтвердите импорт.

Результат: Образ появляется в списке образов соответствующего типа в состоянии «Образ».

При ошибке: Если репозиторий недоступен, импорт завершается с сообщением об ошибке.

Создание контейнера

Предварительные условия: Образ импортирован, настроены профили MIG (при необходимости).

Шаги:

1. В Каталоге выберите образ.
2. Нажмите «Запустить» или «Создать контейнер».
3. Заполните параметры:
   - Хост — выбор из доступных хостов
   - Видеокарта — выбор GPU (отображаются только карты на выбранном хосте)
   - Профиль MIG — если выбранная видеокарта разделена на профили
   - Группы пользователей — для кого доступен контейнер
4. Нажмите «Создать и запустить».

Проверки системы:

- Анализ доступных ресурсов на выбранном хосте
- При недостатке ресурсов — сообщение об ошибке, контейнер не создается
- После успешного запуска контейнер отображается в списке и становится доступен в Чате и Холсте

Особенности создания контейнера с LLM

Дополнительно указывается:

- API-ключ (если требуется)

Автоматически копируются из образа:

- Инференс-движок
- Источник
- Лицензия
- Семейство
- Размерность
- Температура, Top P
- Максимальный размер ответа
- Размер контекстного окна

Остановка и запуск контейнера

Шаги:

1. В Каталоге выберите запущенный контейнер.
2. Нажмите «Остановить».

3. Система освобождает ресурсы, занятые контейнером.

4. Для повторного запуска выберите контейнер и нажмите «Запустить».

5. При запуске можно изменить хост, видеокарту и профиль MIG (аналогично созданию).

Ограничение: Для изменения профилей MIG требуется полное удаление контейнера.

Удаление образа

Шаги:

1. В Каталоге выберите образ (в состоянии «Образ»).

2. Нажмите «Удалить».

3. Подтвердите удаление.

Результат: Образ полностью удаляется из каталога, информация о нем не отображается.

Удаление контейнера

Ограничение: Удалить можно только остановленный контейнер.

Шаги:

1. Остановите контейнер (если он запущен).

2. Выберите контейнер в списке.

3. Нажмите «Удалить».

4. Подтвердите удаление.

Результат: Контейнер полностью удаляется, не отображается в Каталоге и приложениях.

Годовикова 9с17
7 495 234 2222
info@skala-

Чат

Чат — основной пользовательский интерфейс для взаимодействия с LLM-моделями и AI-приложениями.

Выбор модели

Действие**:** В верхней части чата выберите модель из выпадающего списка.

Отображается: Только запущенные контейнеры с LLM-моделями.

Если модель не выбрана: При отправке запроса система возвращает сообщение «Модель не выбрана».

Формирование и отправка вопроса

Шаги:

1. Введите вопрос в поле ввода (поддерживается многострочный ввод).
2. При необходимости прикрепите документы (для RAG).
3. Нажмите «Отправить» или клавишу Enter.

Индикация:

- Во время формирования ответа отображается индикатор «Система думает...»
- Если модель поддерживает стриминг, ответ отображается последовательно, по мере генерации

Управление диалогом

- Новый чат — очищает текущий контекст диалога
- Копировать — копирует вопрос или ответ в буфер обмена

- История — просмотр предыдущих диалогов (фильтрация по периоду: «На этой неделе» / «В этом месяце»)

Фильтрация промтов (Контент-файервол)

Доступно: При запущенном контейнере с приложением фильтрации

Включение/отключение:

1. В боковой панели найдите переключатель «Фильтрация промтов».
2. Включите или отключите фильтрацию.

При включенной фильтрации:

- Запросы проверяются на наличие запрещенного контента и PII
- При обнаружении — запрос блокируется, выводится соответствующее сообщение
- Ответы также фильтруются

При недоступности сервиса: Выводится сообщение «Сервис фильтрации промптов недоступен», запрос не отправляется.

Подключение AI-приложений (RAG, ADA)

Доступно: При запущенных контейнерах соответствующих приложений

RAG (работа с документами):

1. Включите RAG в боковой панели.
2. Прикрепите документы (PDF, DOCX, TXT, изображения).
3. Задайте вопрос по содержанию документов.

4. Система обработает документы (может занять время) и выдаст ответ с указанием источников.

ADA (агентская платформа):

1. Включите ADA в боковой панели.

2. Опишите многоступенчатую задачу на естественном языке.

3. ADA построит план, распределит подзадачи, выполнит их и суммаризирует результат.

При недоступности приложения: Выводится сообщение «Сервис недоступен, попробуйте позже».

OCR и транскрибация

OCR (распознавание текста):

1. В Чате прикрепите изображение или документ с изображениями.

2. Система распознает текст и возвращает результат.

Транскрибация (аудио в текст):

1. В чате прикрепите аудиофайл или укажите ссылку на источник.

2. Система выполнит транскрибацию (требуется время, отображается индикация).

3. Результат отображается с временными метками (при поддержке).

Холст

Холст — графический конструктор для визуального проектирования сценариев работы AI-агентов.

Предварительные условия:

- Запущен контейнер с приложением «Холст»
- Запущен контейнер с приложением ADA
- Запущены контейнеры с LLM и другими используемыми сервисами

Создание сценария

Шаги:

1. Откройте «Холст».
2. Нажмите «Создать сценарий».
3. Укажите имя и описание сценария.
4. Перетащите функциональные блоки (агенты, модели, сервисы) из палитры на рабочее поле.
5. Соедините блоки стрелками, формируя логическую цепочку.
6. Нажмите «Сохранить» и «Активировать».

Формат хранения: Сценарий сохраняется в формате YAML или JSON.

Редактирование сценария

Шаги:

1. Выберите существующий сценарий.
2. Внесите изменения (добавление/удаление блоков, изменение соединений).

3. Сохраните изменения.

4. При необходимости активируйте отредактированную версию.

Копирование сценария

Шаги:

1. Выберите существующий сценарий.

2. Нажмите «Копировать».

3. Укажите новое имя.

4. Новый сценарий сохраняется как неактивный. При необходимости активируйте его.

Деактивация и удаление сценария

- Деактивация: Сценарий перестает использоваться ADA, но остается в системе.
- Удаление: Сценарий полностью удаляется из Холста.

Использование в чате

При поступлении запроса в Чат ADA анализирует сохраненные активные сценарии. При совпадении запроса с описанием или шаблоном сценария используется готовый сценарий для выполнения задачи.

Управление ресурсами GPU

Режимы использования видеокарт

Режим	Описание
Exclusive	Видеокарта используется полностью одним контейнером

Режим	Описание
Shared	Ресурсы распределяются между несколькими контейнерами
Time Slicing	(В следующих версиях)

Разделение ресурсов с помощью MIG (Multi-Instance GPU)

Поддерживается: Для NVIDIA GPU с поддержкой MIG

Цель: Гарантированное выделение ресурсов под контейнер. Каждый профиль MIG может быть использован только одним контейнером. Максимальное количество профилей — 7 (зависит от модели видеокарты).

Шаги настройки:

1. Перейдите в раздел «Управление ресурсами GPU».
2. Выберите хост из списка.
3. Система отображает доступные видеокарты.
4. Выберите видеокарту и установите режим использования «Shared».
5. Активируйте настройку разделения.
6. Задайте количество профилей и ресурсы для каждого (вычислительные ядра, память).
7. Сохраните настройки.

Результат: при создании контейнеров вместо целой видеокарты отображаются имена профилей MIG.

Годовикова 9с17
7 495 234 2222
info@skala-

Перераспределение ресурсов

Ограничение: для изменения конфигурации MIG необходимо остановить и удалить все контейнеры, использующие видеокарту.

Шаги:

1. В разделе «Управление ресурсами GPU» выберите видеокарту.
2. Просмотрите список контейнеров, использующих эту видеокарту.
3. Остановите и удалите все эти контейнеры.
4. Измените количество и размер профилей.
5. Сохраните настройки.

Мониторинг

Основные метрики

Инфраструктура:

- Статус хостов (Готов / Недоступен)
- Загрузка CPU, RAM, GPU
- Дисковое пространство
- Список GPU-карт и MIG-профилей с нагрузкой

Контейнеры:

- Список запущенных контейнеров
- Выделенные ресурсы (CPU, Memory, GPU)

- Время создания
- Кто создал

Использование LLM (в разрезе времени, пользователя, группы, модели):

- Количество входных токенов
- Количество выходных токенов

Журнал аудита

Фиксируемые события:

- Вход в систему
- Управление пользователями (создание, удаление, изменение прав)
- Изменение параметров и системных настроек
- Изменения объектов (создание, запуск, остановка, удаление контейнеров)
- Действия пользователей, влияющие на безопасность

Формат: Фиксируется «было — стало» с указанием автора действия. Журнал недоступен для редактирования.

Настройки приложения

Управление пользователями

Создание учетной записи

Шаги:

1. Перейдите в «Настройки» → «Пользователи».

2. Нажмите «Создать пользователя».

3. Заполните поля:

- Логин — уникальное имя для входа
- Пароль — минимум 8 символов
- Роль — Администратор / MLOps / Пользователь / Аудитор
- Группы — выбор из существующих групп (по умолчанию «default»)
- Статус — Активен / Неактивен

4. Нажмите «Сохранить».

Создание группы

Шаги:

1. Перейдите в «Настройки» → «Группы».

2. Нажмите «Создать группу».

3. Заполните поля:

- Название группы
- Описание (необязательно)
- Роли в группе — мультивыбор (Администратор, MLOps, Пользователь)
- Область доступа — вся платформа / определенные проекты / определенные ресурсы
- Учетные записи — добавление пользователей в группу

4. Нажмите «Сохранить».

Удаление учетной записи

1. В списке пользователей выберите учетную запись.
2. Нажмите «Удалить».
3. Подтвердите удаление.

Управление репозиториями

Docker Registry

Цель: Хранение и распространение образов контейнеров.

Регистрация:

1. Перейдите в «Настройки» → «Docker registry».
2. Нажмите «Добавить репозиторий».
3. Укажите:
   - URL (например, https://registry.skala-ai.ru)
   - Тип — публичный / частный
   - Учетные данные (для частных репозиториев)
4. Нажмите «Сохранить».

Изменение: Выберите репозиторий из списка, отредактируйте параметры, сохраните.

Репозиторий для LLM

Регистрация репозитория:

1. Перейдите в «Настройки» → «Репозитории LLM».

2. Нажмите «Добавить».

3. Укажите:

- URL репозитория
- Токен доступа (при необходимости)

4. Нажмите «Сохранить».