ООО «Апэрбот»
ОГРН 1195081074924 · ИНН 5032312072
115114, г. Москва, ул. Летниковская, д. 10, стр. 4, помещ. 1П
E-mail: team@aprbot.com · Тел.: +7 916 633-03-67

Руководство по эксплуатации программного обеспечения «AIMOS»

1. Вход и пользователи

Вход в веб-интерфейс выполняется по адресу развернутого экземпляра. Администратор создает пользователей через Django admin или CLI (./manage.sh users), назначает роли и права. Программный доступ приложений осуществляется через API-ключи.

2. Управление моделями

- Создание модели: указать workspace/name, тип (llm, vlm, stt, tts, embedder, reranker, ocr, guardrail, processor, subscription, indexer), loader и конфигурацию;
- Загрузка артефактов локально или из Hugging Face;
- Запуск: выбор устройства и порта, старт runtime (vLLM/Ollama/Transformers/специализированный loader), health check;
- Остановка и освобождение ресурсов (GPU/VRAM);
- Монитор отображает фактические состояния (draft/ready/running/stopped/error).

3. API-ключи

В разделе ключей создается ключ AIMOS (sk-...), связывается с одной или несколькими моделями, задаются алиасы, модель по умолчанию, лимиты входных/выходных/суммарных токенов и период действия. Ключ можно активировать и отозвать. Использование и стоимость накапливаются на ключе и модели.

4. Работа с задачами и Actions

Длительные операции (обработка аудио, документов, изображений) отображаются как задачи со статусами pending/running/completed/failed/cancelled, приоритетом, прогрессом и выходными файлами. Actions предоставляют OCR, конвертацию, извлечение текста/изображений, транскрипцию и иные processor-действия по динамически формируемой схеме.

5. RAG и индексы

Для работы с корпоративной базой знаний создается подключение к indexer-сервису, загружаются документы, выполняется семантический поиск через веб-интерфейс или клиентский /v1/indexer API.

6. Анонимизация данных (guardrail-модель)

При подключенной guardrail-модели пользовательские сообщения перед передачей во внешние модели проходят анонимизацию персональных и конфиденциальных данных; ответы модели деанонимизируются перед возвратом клиенту.

7. Мониторинг, резервирование, обновление

- Наблюдаемость: health endpoints, метрики, логи моделей с выгрузкой, OpenTelemetry, Prometheus, опциональный Sentry;
- Резервное копирование: согласованный снимок базы данных и файловых директорий (data/workspaces, data/media, data/tasks, data/logs, data/envs, data/venvs);
- Обновление: получение обновленного кода из репозитория, make migrate, make build, контролируемый перезапуск сервисов;
- Диагностика: просмотр логов сервисов и моделей, статусов процессов, портов и GPU.

Исх. № РПО-2026/10-16/АГ-06

16 октября 2026 г.

Генеральный директор ООО «Апэрбот»



/ А б о л м а с о в А . Г . /

м.п.