**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

# Функциональные характеристики программного обеспечения «AIMOS»

## 1. Назначение

AIMOS — сервер и панель управления AI-моделями. Программное обеспечение объединяет локальные модели, внешние AI API и подписочные аккаунты в единую управляемую среду с общими ключами доступа, совместимыми API, учётом использования, очередями задач и веб-интерфейсом.

## 2. Роли пользователей

| | |
|---|---|
| **Администратор** | Пользователи и права, ключи, лимиты, видимость моделей, инфраструктура |
| **ML-инженер** | Регистрация и запуск локальных моделей, управление жизненным циклом |
| **Разработчик** | Подключение приложений через SDK и единый base URL |
| **Оператор/пользователь** | Работа с задачами, документами, индексами через веб-интерфейс |

## 3. Основные функции

### Каталог моделей

Хранение моделей в пространстве имён workspace/name; типы LLM, VLM, embedder, STT, TTS, reranker, guardrail, processor, subscription, indexer, OCR; состояния draft/ready/running/stopped/error; цены и видимость

### Локальный запуск моделей

Загрузчики vLLM, Ollama, SGLang, Transformers и специализированные; загрузка артефактов из Hugging Face; выбор устройства и порта, память, tensor parallelism; фоновый монитор состояния

### Унифицированный AI API

OpenAI Chat Completions, Completions, Responses, Embeddings, Files, Images, Audio; Anthropic Messages; нативные маршруты Ollama и Kimi; потоковые ответы

### API-ключи и использование

Ключ на одну или несколько моделей, алиасы, активация/отзыв, лимиты токенов, сроки действия, счётчики, расчёт стоимости

### Подписки и AI-клиенты

Шлюз к subscription-моделям с OAuth-управлением токенов; поддержка Codex CLI, Claude Code, Cursor, OpenAI/Anthropic SDK

**Задачи и очереди**

Синхронные и асинхронные операции со статусами, приоритетом, прогрессом, временем ожидания/выполнения и выходными файлами

**RAG и индексация**

Подключение индексов, загрузка и удаление документов, семантический поиск, клиентский /v1/indexer API

**Анонимизация данных (guardrail-модель)**

Анонимизация персональных данных в запросах перед внешними моделями и деанонимизация ответов

**Веб-интерфейс**

Каталог и карточки моделей, создание моделей и подписок, ключи, пользователи, задачи, обучение, Actions, индексы, изображения, ресурсы оборудования, API-документация; RU/EN

**Наблюдаемость**

Health endpoints, метрики системы и моделей, логи с выгрузкой, OpenTelemetry, Prometheus, опциональный Sentry

**4. Ограничения**

- GPU/CUDA требуется для соответствующих загрузчиков; CPU-режим доступен не для всех моделей;
- совместимость API означает совместимость формата реализованных маршрутов, а не полную идентичность облачным API;
- сторонние внешние сервисы (providers, processors, indexer, guardrail-модель) подключаются опционально и не входят в поставку.

Исх. № РПО-2026/10-16/АГ-04

16 октября 2026 г.

Генеральный директор ООО «Апэрбот»



/ А б о л м а с о в  А . Г . /

м.п.