AI BOX

ОПИСАНИЕ ФУНКЦИОНАЛЬНЫХ ХАРАКТЕРИСТИК

Версия 2025.3.0

powered by IT Expert

СОДЕРЖАНИЕ

ТЕРМИНЫ И СОКРАЩЕНИЯ

- **LLM –** large Language Model, большая языковая модель.
- **АРМ –** автоматизированное рабочее место.
- **СУБД** – система управления базами данных.
- **Токен –** минимальная единица текста, с которой работает LLM.

1 НАЗНАЧЕНИЕ ПРОДУКТА

Программный продукт AI BOX - это корпоративный AI-продукт, обеспечивающий безопасное подключение, администрирование и мониторинг LLM и моделей эмбеддингов, а также интеграцию с корпоративными источниками знаний и внешними сервисами.

AI BOX объединяет административные функции, аналитические инструменты и API, создавая инфраструктурный слой для построения и масштабирования интеллектуальных решений в корпоративной среде.

AI BOX предназначен для:

– интеграции LLM и моделей эмбеддингов в корпоративную инфраструктуру;

– построения широкого спектра интеллектуальных приложений, включая чат-ботов для поддержки клиентов и сотрудников, извлечение структурированных и неструктурированных данных, семантический анализ текста, доступ к корпоративным данным на естественном языке, формирование аналитических дайджестов и т.д.

2 ОПИСАНИЕ ФУНКЦИОНАЛЬНЫХ ХАРАКТЕРИСТИК

2.1 Управление подключениями к моделям

Настройка подключений к внешним провайдерам LLM (Ggigachat, Яндекс и т.д.). Настройка подключений к LLM, развернутым в закрытом контуре. Управление ключами и параметрами API. Регистрация LLM и моделей эмбеддингов, настройка тарифов токенов и их назначения (генерация или индексация).

Таким образом, AI BOX выступает как корпоративный шлюз, обеспечивающий единый API-доступ к разным LLM-провайдерам (внутренним и внешним).

2.2 Управление промптами

Разработка, редактирование и тестирование шаблонов промптов для различных задач.

2.3 Управление областями знаний

Создание, загрузка и индексация корпоративных источников данных. Поддержка поиска и тестирования ответов чат-бота.

Настройка коннекторов, загрузка данных в области знаний

2.4 Мониторинг использования

Визуализация статистики по расходу токенов, количеству запросов, стоимости и активности моделей.

2.5 Безопасность и управление доступом

AI BOX обеспечивает высокий уровень безопасности данных и взаимодействия:

- аутентификация через Keycloak (OIDC, OAuth2);
- создание, блокировка, удаление и восстановление учётных записей пользователей и сервисов;
- ролевая модель доступа (администратор, пользователь, интеграция);
- шифрование API-ключей и токенов;
- аудит всех действий в административной консоли и API.

2.6 Возможности API и интеграции

AI BOX предоставляет REST API-интерфейс для любых корпоративных приложений.

Основные функции API:

- выполнение текстовых запросов к LLM (генерация, анализ, классификация);
- формирование векторных представлений и поиск по контенту;
- управление диалогом с пользователем в режиме чата (API для AI ассистента);
- работа с шаблонами промптов;
- поддержка журналирования и аудита запросов;
- работа в синхронном и асинхронном режимах.

3 ОСНОВНЫЕ ТЕХНИЧЕСКИЕ ХАРАКТЕРИСТИКИ

3.1 Надежность

Надежность функций, реализованных в программном обеспечении AI BOX, обеспечивается следующим комплексом мер:

– встраивание в алгоритмы реализации функций обработчиков исключительных ситуаций;

– аварийное завершение AI BOX не приводит к отказу связанных смежных систем;

– задачи AI BOX реализованы так, что действия пользователей не приводят к сбоям программного обеспечения, либо к аварийному завершению.

Надежность систем на базе программного обеспечения AI BOX обеспечивается следующим комплексом мер:

– использование лицензионного программного обеспечения;

– использование отказоустойчивой СУБД с возможностью восстановления данных после сбоя;

– контроль за целостностью данных на уровне СУБД;

– использование отказоустойчивых компонент серверного и телекоммуникационного оборудования;

– защита серверов и телекоммуникационного оборудования от сбоев в электропитании, достигаемая за счет дублирования энергоснабжения и системы аварийного переключения резервного электроснабжения;

– применение систем бесперебойного электропитания технических средств со временем автономной работы, достаточным для принятия необходимых мер по сохранению всех данных и корректной остановки AI BOX при возникновении неполадок в энергоснабжении;

– использование средств резервного копирования и восстановления данных.

3.2 Масштабируемость

Масштабируемость систем на базе AI BOX обеспечивается по следующим параметрам:

– количеству пользователей;

– количеству одновременно работающих пользователей;

– количеству обрабатываемой информации.

Масштабируемость обеспечивается без модификации программного обеспечения путём:

– добавления дополнительных серверных мощностей;

– применения мульти серверной архитектуры (т.е. путём организации кластеров СУБД, сервера приложений и/или веб-сервера).

3.3 Доступность и производительность

Типовые параметры доступности и производительности для систем на базе AI BOX приведены в таблице 1.

Таблица 1. Параметры доступности и производительности

Показатель	Значение
Штатный режим работы (период доступности системы) с учетом технологических перерывов на проведение регламентных и профилактических работ	24 часа, технологический перерыв 2 часа
Критичные для выполнения бизнес-функций периоды функционирования системы	Рабочее время
Максимальное время недоступности системы (ее компонент) / Максимальное время, отведенного для восстановления системы (целевое время восстановления)	1 рабочий день
Регламентные (зарезервированные) периоды проведения плановых работ, не требующих получения разрешения на их проведение	1 раз неделю в течение 2-х часов
Время отклика системы на запрос ресурса (без учета влияния сетевой инфраструктуры)	• время входа в систему зарегистрированного пользователя Системы – не более 3 сек.; • время открытия формы объекта в Системе – не более 3 сек.; • время открытия списка объектов Системы (реестра) – не более 5 сек.; • время операций поиска данных – не более 15 сек.; • время формирования отчета (без подотчетов) – не более 20 сек. (на этапе проектирования должны быть разработаны технические и организационные мероприятия, проводимые в случае возникновения задержек и замедления работы системы при формировании отчетов).

4 РЕСУРСЫ, НЕОБХОДИМЫЕ ДЛЯ РАБОТЫ

AI BOX функционирует в архитектуре «клиент-сервер». Доступ к функциям AI BOX осуществляется через автоматизированные рабочие места (АРМ), объединенные локальной вычислительной сетью.

Установки клиентской части AI BOX на АРМ не требуется. Доступ к функциям осуществляется с использованием web-браузера. Рекомендуемые требования к характеристикам АРМ приведены в таблице 2.

Таблица 2. Минимальные требования к характеристикам АРМ

Параметр	Рекомендуемое значение
Разрешение монитора	1366 x 768 или FHD 1920 x 1080 16:9
Процессор	Intel Core-i3
Оперативная память	4GB
SSD / HDD	256 GB
Операционная система	Linux Windows
Web-браузер	Яндекс Браузер Google Chrome, Mozilla Firefox, Microsoft Edge, Apple Safari
Скорость обмена данными между клиентом и сервером	1 Мбит/с