ФИЛОСОФИЯ.ИТ
РОСАТОМ

Общество с ограниченной ответственностью «Философия.ИТ»

Россия, 107023, г. Москва, ул. Измайловский Вал, д. 30
Тел.: +7 (495) 988-37-38, факс: +7 (495) 988-37-38,
e-mail: customer@fil-it.ru

ИНН 7713728490
КПП 771901001
ОГРН 1117746379145

ОПИСАНИЕ ТЕХНИЧЕСКОЙ АРХИТЕКТУРЫ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ

ПЛАТФОРМА ИСКУССТВЕННОГО ИНТЕЛЛЕКТА «КОГНИТРОН»

г. Москва

Оглавление

1. Общие положения 3
1.1. Полное наименование программы для ЭВМ, обозначение 3
1.2. Назначение системы 3
1.3. Разработчик системы 3
1.4. Назначение документа 3
2. Технологический стек 4
3. Техническая архитектура 5
4. Взаимодействие компонентов 8

1. Общие положения

1.1. Полное наименование программы для ЭВМ, обозначение

Полное наименование Программы для ЭВМ: Платформа искусственного интеллекта «Когнитрон» – далее по тексту Система.

1.2. Назначение системы

«Когнитрон» - комплексное программное решение, объединяющее инструменты обучения и развертывания моделей искусственного интеллекта и создания ИИ-помощников и ИИ-агентов. Платформа поддерживает работу с большими языковыми моделями (LLM) модальности любого типа: текст, изображение, аудио, видео. Решение выступает в роли среды, обеспечивающей коммуникацию и координацию ИИ-агентов между собой, и средства интеграции с внешними ИТ-системами.

1.3. Разработчик системы

Полное наименование: Общество с ограниченной ответственностью «Философия.ИТ». Сокращенное наименование: ООО «Философия.ИТ».

1.4. Назначение документа

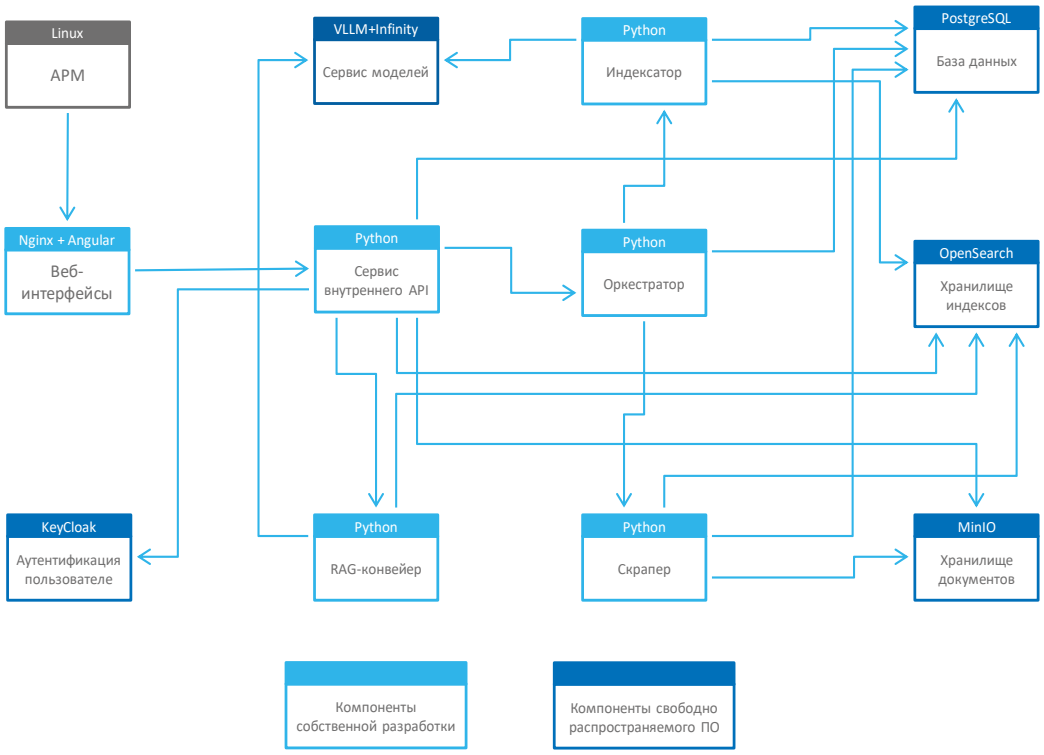
Настоящий документ входит в комплект эксплуатационной документации по Платформе искусственного интеллекта «Когнитрон» и содержит описание технической архитектуры Системы.

2. Технологический стек

Компонент	Описание
Python	Высокоуровневый, интерпретируемый язык программирования, поддерживает объектно-ориентированное, процедурное и функциональное программирование.
FastAPI	Веб-фреймворк на Python для создания API с использованием стандартных аннотаций типов.
LangChain	Фреймворк модульных компонент для создания цепочек обработки данных, интеграции с API, базами данных и другими инструментами для объединения LLM с внешними источниками данных и сервисами.
TypeScript	Язык программирования, расширяющий JavaScript за счет статической типизации и дополнительных возможностей
Angular	Фреймворк создания одностраничных веб-приложений (SPA) на TypeScript.
OpenSearch	Поисковая и аналитическая система, с расширенными возможностями для работы с текстовой и векторной информацией для систем семантического поиска и приложений на основе искусственного интеллекта.
PosrgreSQL	Объектно-реляционная система управления базами данных с открытым исходным кодом, поддерживает SQL- и JSON-запросы, используется для хранения технической информации; хранение бизнес-информации в данной БД не предусмотрено.
VLLM	Библиотека обслуживания больших языковых моделей (LLM).
Infinity	Фреймворк векторного поиска, совместимый с Sentence Transformers и другими embedding-моделями.
Sentence-Transformers	Библиотека для получения векторных представлений предложений и текстов, основанная на модели BERT и её производных. Используется для семантического поиска и кластеризации текста.
Transformers	Библиотека Hugging Face для работы с современными моделями NLP, CV и мультимодальными моделями.

Компонент	Описание
Unstructured	Инструмент извлечения и структурирования данных из неструктурированных документов (PDF, DOCX, HTML и т.п.) для дальнейшего анализа или индексации.
Celery	Распределённая система управления очередями задач, позволяющая выполнять фоновые и периодические задачи в Python.
Redis	Хранилище данных в памяти для кэширования и брокера сообщений.
NGINX	Веб-сервер и обратный прокси.
MinIO	Объектное хранилище, совместимое с Amazon S3.
Keycloak	Система управления идентификацией и доступом (IAM) с поддержкой SSO, поддерживает протоколы OpenID Connect, OAuth 2.0 и SAML.

3. Техническая архитектура



Компоненты собственной разработки

Сервис внутреннего API – Центральный точка входа для взаимодействия между компонентами системы. Предоставляет RESTful интерфейс для управления данными, запуска процессов и получения результатов. Обеспечивает маршрутизацию запросов, аутентификацию и логирование операций.

RAG-конвейер – Система Retrieval-Augmented Generation для обогащения запросов релевантной информацией из хранилищ документов. Осуществляет поиск контекста, ранжирование результатов и передачу дополнительных данных в языковую модель для более точных ответов.

Оркестратор – Управляет потоком выполнения рабочих процессов наполнения хранилища документов и хранилища индексов. Координирует взаимодействие между сервисами, обеспечивает обработку ошибок, повторные попытки и отслеживание статуса выполнения задач.

Скрапер – Автоматизирует сбор данных из внешних источников (веб-сайты, API, документы). Выполняет парсинг, нормализацию и первичную обработку информации для последующего индексирования и использования в системе.

Индексатор – Преобразует неструктурированные данные в индексированный формат, оптимизированный для быстрого поиска. Создает семантические представления, применяет NLP-преобразования и сохраняет структурированную информацию в поисковом индексе.

Web-интерфейс пользователя – Приложение для конечных пользователей, предоставляющее интуитивный доступ к функциям поиска, анализа и взаимодействия с системой. Позволяет выполнять запросы к RAG-конвейеру, просматривать результаты и управлять персональными настройками.

Web-интерфейс администратора – Выделенная панель управления для администраторов и операторов системы. Обеспечивает мониторинг состояния компонентов, управление пользователями, настройку параметров индексации, просмотр логов и аналитики работы системы.

Компоненты свободно распространяемого ПО

Сервис моделей (vLLM, Infinity) – Управляет жизненным циклом машинных моделей: их загрузкой, развертыванием и инференсом. Обеспечивает API для обращения к различным моделям (LLM, embedding-модели), кэширование и оптимизацию производительности.

Сервис аутентификация и авторизация (Keycloak) – Обеспечивает безопасное управление доступом к ресурсам системы. Выполняет проверку учетных данных, выдачу токенов, управление ролями и разрешениями для пользователей и приложений.

База данных (PostgreSQL) – Реляционная база данных для хранения метаинформации, конфигурации, результаты обработки, логи операций.

Хранилище индексов (Opensearch) – Специализированное хранилище для семантических индексов и векторных представлений для поиска по сходству, поддерживает векторные операции и фильтрацию по метаданным.

Хранилище документов (MinIO) – Долгосрочное хранилище исходных документов и необработанных данных. Обеспечивает сохранение, версионирование, быстрый доступ к оригинальным материалам для воспроизводимости.

4. Взаимодействие компонентов

Пользователи подключаются к системе по протоколу HTTP(S) используя веб-браузер. Компоненты системы взаимодействуют по сетевому протоколу TCP в соответствии со спецификациями.