



Общество с ограниченной ответственностью «Скайори»
шоссе Космонавтов., д.111И к1, помещ. 77, Пермь, 614066
ОГРН 1195958032181, ИНН 5905061802, КПП 590501001

Корпоративный ИИ Эксперт «Skyori AI Expert»

Описание технической архитектуры

Листов 7

Пермь, 2026

общие сведения

Корпоративный ИИ Эксперт «Skyori AI Expert» (далее – ИИ Эксперт, система) является функциональным решением, предназначенным для автоматизации процессов анализа документов и предоставления информационно-аналитических материалов пользователям. Система обеспечивает оперативный поиск релевантной информации в массиве документов по заданным тематическим вопросам, автоматизацию подготовки аналитических материалов с указанием использованных источников, а также диалоговый режим работы для уточнения запросов и поддержания контекста беседы.

ИИ Эксперт предназначен для решения следующих задач:

- сбор, обработка и систематизация текстовых документов;
- обеспечение семантического поиска релевантных фрагментов документов с использованием технологий векторного поиска и больших языковых моделей;
- автоматизация подготовки аналитических материалов, содержащих обобщенную информацию по загруженным документам с указанием источников;
- обеспечение диалогового режима взаимодействия с системой для уточнения запросов и поддержания контекста беседы;
- снижение временных затрат сотрудников на рутинные операции по поиску и первичному анализу текстовой информации.

ИИ Эксперт реализован в виде веб-приложения.

описание технической архитектуры

структурные элементы и их интерфейсы, с помощью которых составлена система

В структурные элементы, с помощью которых составлена система, входят объекты метаданных и технологические элементы.

Объекты метаданных: массивы данных, документы, отчеты.

В ИИ Эксперте ведутся следующие массивы данных:

- текстовые документы, загружаемые в объектное хранилище;
- индексы – наборы документов, организованные в иерархическую структуру (вершины и листья);
- фрагменты документов – результат разбиения исходных документов на смысловые части для последующего векторного поиска;
- векторные представления (эмбединги) текстовых фрагментов и пользовательских запросов;
- сообщения пользователей и ответы системы (история чатов);
- настройки чатов: параметры генерации ответов, выбранные индексы, системные промпты;
- пользователи и роли доступа;

- статусы индексации: Ready, InProcess, HasErrors;
 - метаданные документов: названия, даты, привязка к индексам.
- технологические элементы: режимы функционирования, выполняемые функции, экранные интерфейсы.

Режимы функционирования:

- штатный режим функционирования;
- аварийный режим функционирования.

Основным режимом функционирования системы является штатный режим. В штатном режиме функционирования обеспечивается:

- клиентское программное обеспечение и технические средства пользователей и администратора ИИ Эксперта обеспечивают возможность функционирования в круглосуточном режиме;
- серверное программное обеспечение и технические средства серверов обеспечивают возможность круглосуточного функционирования, с перерывами на обслуживание;
- обеспечивается исправная работа оборудования, составляющего комплекс технических средств;
- обеспечивается исправное функционирование системного, базового и прикладного программного обеспечения ИИ Эксперта.

Для обеспечения штатного режима функционирования системы Заказчик выполняет требования и выдерживает условия эксплуатации программного обеспечения и комплекса технических средств, указанные в соответствующих технических документах (техническая документация, инструкции по эксплуатации и т.д.).

Аварийный режим функционирования системы характеризуется отказом одного или нескольких компонент программного и (или) технического обеспечения. В случае перехода системы в предаварийный режим необходимо предусмотреть:

- завершение работы всех приложений, с сохранением данных;
- выполнение резервного копирования базы данных (далее – БД) и объектного хранилища документов.

После этого необходимо выполнить комплекс мероприятий по устранению причины перехода системы в аварийный режим.

Выполняемые системой функции:

- загрузка и хранение текстовых документов в объектном хранилище с организацией иерархической структуры индексов;
- автоматическая индексация загруженных документов: извлечение текста, разбиение на фрагменты, векторизация с помощью модели эмбедингов и сохранение в векторном хранилище для последующего семантического поиска;

- семантический поиск релевантных фрагментов документов по запросу пользователя на основе векторных представлений (эмбедингов);
- генерация структурированных ответов на запросы пользователей с использованием больших языковых моделей на основе найденных релевантных фрагментов документов и контекста диалога;
- отображение блока источников с указанием использованных документов и коэффициента релевантности каждого источника, с возможностью просмотра выдержки из текста;
- диалоговый режим работы с поддержанием контекста беседы (передача предыдущих запросов и ответов на вход нового запроса для уточнения и детализации);
- одновременный запрос к нескольким индексам с компоновкой результатов в единый ответ;
- управление чатами (создание, переименование, удаление, добавление в избранное, поиск по списку чатов);
- управление источниками поиска (добавление, удаление документов, участвующих в генерации ответа);
- экспорт результатов запросов и перечня использованных источников в текстовом формате;
- настройка параметров генерации ответов: выбор языковой модели, количества источников, температуры и других параметров;
- управление промптами – создание и редактирование системных инструкций для настройки поведения языковой модели при генерации ответов;
- управление индексами: создание, редактирование, удаление индексов и вершин иерархии, запуск индексации, мониторинг статусов индексации.

Ролевой моделью ИИ Эксперта определены автоматизированные рабочие места следующих пользователей: пользователь, администратор.

Пользователю предоставляются следующие возможности:

- создание новых чатов и отправка запросов;
- выбор индексов и настроек чата для поиска;
- просмотр истории сообщений в формате «вопрос-ответ» с форматированными ответами и блоком источников;
- работа с чатами: переименование, удаление, добавление в избранное, поиск по списку чатов;
- работа с источниками: добавление, удаление источников поиска;
- экспорт чата и источников в текстовом формате.

Администратор обладает всеми правами пользователя, а также дополнительно имеет следующие возможности:

- управление индексами (создание, редактирование, удаление индексов и вершин иерархии);
- запуск индексации всех индексов или отдельного индекса;
- управление настройками чатов: создание и редактирование промптов, настройка параметров генерации ответов (языковая модель, количество источников, количество фрагментов, размер контекста, температура).

Интерфейс администратора предоставляет доступ ко всем функциям ИИ Эксперта, включая управление пользователями и ролями через центр администрирования Keuscloak, а также участие в процедурах установки, обновления и мониторинга работоспособности системы.

оединение элементов структуры

Структурные элементы системы на уровне интерфейсов имеют следующие соединения:

- переход к списку чатов из главной страницы системы;
- переход к диалогу с ИИ Экспертом при выборе или создании чата;
- переход к настройкам чата (выбор индексов, параметров генерации) из области диалога;
- переход к странице управления индексами из меню навигации администратора;
- переход к странице создания и редактирования индекса из списка индексов;
- переход к запуску индексации со страницы редактирования индекса или из списка индексов;
- переход к странице настроек чатов из меню навигации администратора;
- переход к странице создания и редактирования настройки из списка настроек;
- переход к просмотру выдержки текста источника при клике на источник в блоке источников ответа.

Соединение элементов структуры на уровне архитектурных решений описаны в п. 2.3 настоящего документа.

рхитектурный стиль, который направляет все элементы, их интерфейсы, их сотрудничество и их соединение

ИИ Эксперт спроектирован на основе клиент-серверной архитектуры с использованием доступа через веб-приложение (Рисунок 1):

- SkyoriAIExpert – docker-образ, содержащий веб-сервер и веб-приложение ИИ Эксперта:
 - Ngnix – веб-сервер, выполняющий обработку запросов от браузеров пользователей;

- AIExpert – веб-приложение, реализующее экранные формы и обработку запросов пользователей.
- Keycloak – сервис аутентификации и управления доступом
- AIExpert DB – реляционная база данных ИИ Эксперта (PostgreSQL).
- Documents – объектное хранилище для исходных файлов документов (MinIO).
 - поисковый движок для хранения векторных представлений документов и выполнения семантического поиска (OpenSearch).
- LLM – подключение к внешней языковой модели для генерации ответов на запросы пользователей.

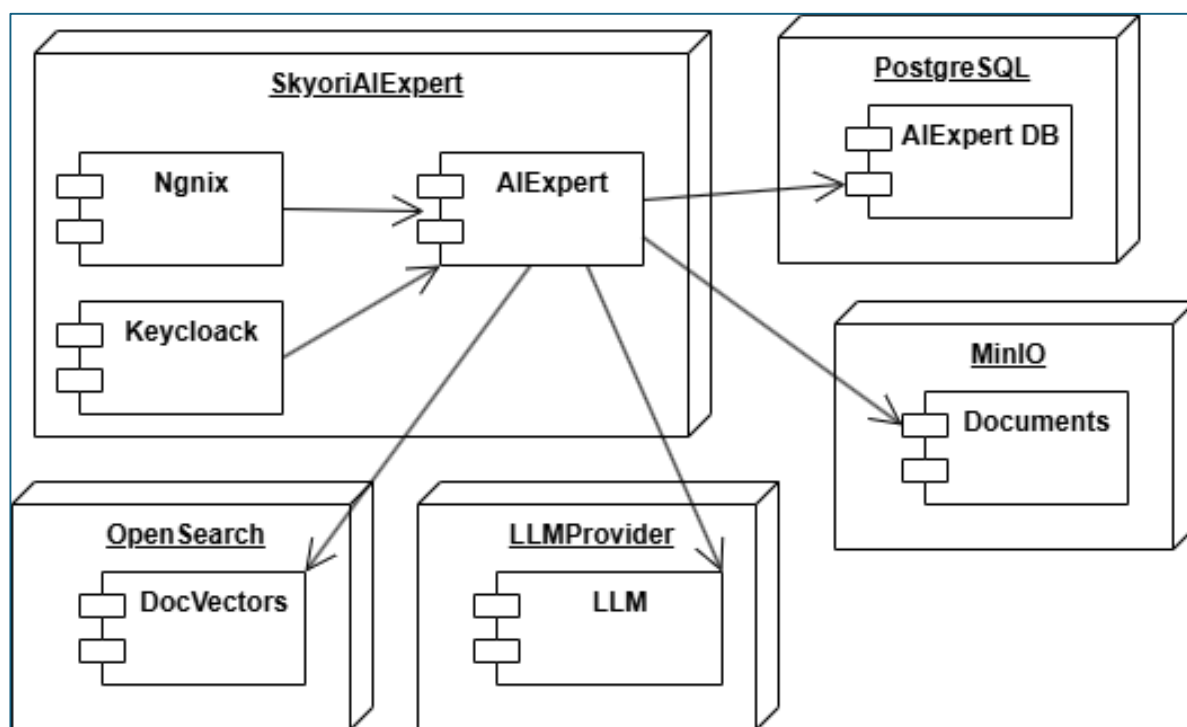


Рисунок 1. Архитектура ИИ Эксперта

Описание соединений элементов архитектуры:

- Соединение «Nginx → AIExpert»: запросы на отображение экранных форм и выполнение операций ИИ Эксперта.
- Соединение «Keycloak → AIExpert»: проверка учетных данных пользователей, получение информации о ролях и правах доступа.
- Соединение «AIExpert → AIExpert DB»: запросы на получение, создание, изменение и удаление данных в базе данных.
- Соединение «AIExpert → Documents»: запросы на загрузку, хранение и получение исходных файлов документов из объектного хранилища MinIO.
- Соединение «AIExpert → DocVectors»: запросы на выполнение семантического поиска по векторным представлениям фрагментов документов в OpenSearch.

- Соединение «AIExpert → LLM»: запросы на отправку сформированных запросов к языковой модели и получение сгенерированных ответов.

ИИ Эксперт организован на основе клиент-серверной архитектуры с использованием доступа через веб-приложение с использованием следующих технологий:

- Технологии: .NET Core, Docker, LLM, клиент-сервер, микросервисы.
- Язык программирования: C#, Python, TypeScript, HTML, CSS.
- Операционная система: Ubuntu, Debian, CentOS, AstraLinux, РедОС, ALT Linux.
- СУБД: PostgreSQL.
- Объектное хранилище: MinIO.
- Векторное хранилище: OpenSearch.
- Языковая модель: подключение к LLM-провайдерам (осуществляется через стандартные API-интерфейсы (REST API /