

**Платформа интеллектуальной обработки обращений, текста,
документов, картинок, голоса без участия человека с применением
технологий генеративного искусственного интеллекта для решения
отраслевых задач «ГЕН.ИИ»**

ОПИСАНИЕ ФУНКЦИОНАЛЬНЫХ ХАРАКТЕРИСТИК И ТРЕБОВАНИЙ

СОДЕРЖАНИЕ

ГЛОССАРИЙ.....	3
1 ОБЩИЕ СВЕДЕНИЯ.....	5
1.1 Полное наименование Системы и условное обозначение.....	5
1.1 Наименование предприятия-разработчика системы.....	5
2 НАЗНАЧЕНИЕ И ЦЕЛИ СОЗДАНИЯ СИСТЕМЫ.....	6
2.1 Назначение Системы.....	6
2.2 Цель создания Системы.....	6
3 ХАРАКТЕРИСТИКА ОБЪЕКТА АВТОМАТИЗАЦИИ.....	8
3.1 Краткие сведения.....	8
3.2 Участники создаваемой Системы.....	8
4 ОПИСАНИЕ ФУНКЦИОНАЛЬНЫХ ХАРАКТЕРИСТИК СИСТЕМЫ.....	9
4.1 Требования к Системе в целом.....	9
4.1.1 Структура и функционированию Системы.....	10
4.1.2 Надежность.....	13
4.1.3 Безопасность.....	14
4.1.4 Эргономика и техническая эстетика.....	14
4.2 Описание функциональных блоков Системы.....	15
4.2.1 Описание функционального блока «Модуль управления базами знаний».....	16
4.2.2 Описание функционального блока «Чат».....	17
4.2.3 Описание функционального блока «Управление пользователями».....	18
4.3 Техническое обеспечение.....	18
4.4 Входные и выходные данные.....	19
4.4.1 Входные данные.....	19
4.4.2 Выходные данные.....	19

ГЛОССАРИЙ

Для достижения общего понимания сокращенных наименований и терминов, используемых в настоящем документе, предлагается использовать следующий список:

Термин	Определение
БД	База данных
БЗ	База знаний (источник данных)
ИИ	Искусственный интеллект
Индексация (базы знаний)	Подготовка загруженных данных с преобразованием для последующей обработки запросов пользователя
ИТ	Информационные технологии
Векторизация	Один из процессов индексации базы знаний при обработке загруженных документов
Векторная БД	Специализированная система хранения,предназначенная для работы с векторными вложениями (числовыми характеристиками неструктурированных данных)
Мб	Мегабайт - единица измерения количества информации
Микросервис	Веб-сервис, отвечающий за один элемент логики в определенной предметной области
Модуль	Часть системы, отвечающая за определенный набор функций
Модуль интерфейса управления	Веб-интерфейс доступа к управлению Системой
Оператор	Совокупное обозначение пользователей, имеющих ограниченную роль в Системе
ОС	Операционная система
ПО	Программное обеспечение
«Ген.ИИ» для автоматизации входящих обращений /ГЕН.ИИ/Система	Платформа интеллектуальной обработки обращений, текста, документов, картинок, голоса без участия человека с применением технологий генеративного искусственного интеллекта для решения отраслевых задач «ГЕН.ИИ»
Пользователь	Сотрудник организации, который работает с системой
Промт (системный)	Настраиваемая инструкция для модели искусственного интеллекта, помогающая наиболее точно сформулировать ответ
Система ролевого доступа	Модуль, отвечающий за заведение, авторизацию пользователей и назначение прав доступа.
СУБД	Системы управления базами данных
Холодное хранилище	Система, предназначенная для хранения загружаемых документов (файлов) в исходном виде
Чанк	Фрагмент или блок данных, который выделяется из более крупного набора данных для обработки
Эмбеddинг (текстовый)	Числовое представление данных, позволяющее моделям искусственного интеллекта анализировать и интерпретировать текст
HTTPS	Расширение протокола HTTP для поддержки

	шифрования в целях повышения безопасности.
LLM	Большая Языковая Модель, тип искусственного интеллекта, который обучен на больших объемах текстовых данных
mTLS	Двухсторонний протокол TLS
RAG (Retrieval-Augmented Generation)	подход в области обработки естественного языка, который сочетает в себе извлечение информации и генерацию текста
RBAC	Политика управления доступом на основе ролей
REST API	Интерфейс, используемый компьютерными системами для безопасного обмена информацией через сеть Интернет
TLS	Протокол защиты транспортного уровня
.NET Core, Docker, Postgres, Linux	Открытые платформы разработки

1 ОБЩИЕ СВЕДЕНИЯ

1.1 Полное наименование Системы и условное обозначение

Полное наименование Системы: «Ген.ИИ» для автоматизации входящих обращений

Условное обозначение, торговый знак: ГЕН.ИИ

1.1 Наименование предприятия-разработчика системы

Разработчик системы: Общество с ограниченной ответственностью «ДжиДиСи Сервисез».

Адрес разработчика: Республика Татарстан, Лаишевский район, с. Усады, ул. Дорожная, д. 42, корпус 2, офис 835.

2 НАЗНАЧЕНИЕ И ЦЕЛИ СОЗДАНИЯ СИСТЕМЫ

2.1 Назначение Системы

Система «Ген.ИИ» для автоматизации входящих обращений» представляет собой платформу для локального развертывания и состоит из двух основных модулей:

- Интеллектуальный диалоговый поиск по базе знаний и файлам компаний
- Обработка и классификация входящих обращений.

Модуль «Интеллектуальный диалоговый поиск по базе знаний и файлам компаний» основанное на использовании семейства больших языковых моделей (LLM) и разработанной RAG-системы (Retrieval-Augmented Generation). Он предназначен для решения задач интеллектуального поиска, формирования ответов и ведения интеллектуального диалога в организациях различных отраслей. Обработывая и анализируя большие объемы текстовой информации, система нацелена решать рутинные, формализуемые и повторяющиеся операции, такие как:

- обработка документов: извлечение данных из документов,
- быстрый доступ к необходимым данным из разрозненных документов и файловых хранилищ,
- извлечение информации на основе контекста запросов пользователей,
- ежедневные шаблонные процессы взаимодействия с клиентами и пользователями,
- уточнение запросов и предоставление дополнительных сведений в режиме реального времени

Таким образом, ГЕН.ИИ автоматизирует рутинные задачи, связанные с поиском, анализом и обработкой информации и, извлекая ключевые данные из документов, представив их в удобном формате, позволяет сотрудникам сосредоточиться на более сложных и творческих задачах.

2.2 Цель создания Системы

Целями создания являются:

- Сокращение человеческих усилий;
- Улучшение качества сервиса;
- Сокращение человеческих ошибок;
- Повышение эффективности поиска;

- Повышение эффективности управления ресурсами и производительностью сотрудников.

3 ХАРАКТЕРИСТИКА ОБЪЕКТА АВТОМАТИЗАЦИИ

3.1 Краткие сведения

ГЕН.ИИ – инструмент автоматизации обработки и анализа текстовых данных, которые могут поступать из различных источников в форматах docx, pdf, txt, баз знаний и файловых хранилищ. Система обеспечивает эффективный доступ к этой информации и производит ее анализ с помощью инновационных моделей искусственного интеллекта. Система позволяет вести интеллектуальный диалог и состоит из следующих функциональных модулей:

- Холодное хранилище для баз знаний
- Модуль текстового эмбединга (векторизации) и RAG
- Система ролевого доступа
- Модуль LLM
- Модуль управления базами знаний
- Модуль интерфейса управления

3.2 Участники создаваемой Системы

Основные роли участников создаваемой Системы приведены в таблице 1.

Таблица 1. Основные роли участников создаваемой Системы

Роль участника	Описание
Администратор	Сотрудник, имеющий полный доступ ко всем функциям, участвует на всех этапах работы с системой и может управлять пользователями, определяя их роли.
Менеджер	Сотрудник, который участвует в процессах, относящихся к управлению бизнес логикой системы, включая управление некоторыми функциями, доступными операторам.
Оператор	Сотрудник, который является конечным (целевым) пользователем системы и участвует исключительно на этапе ведения интеллектуального диалога.

4 ОПИСАНИЕ ФУНКЦИОНАЛЬНЫХ ХАРАКТЕРИСТИК СИСТЕМЫ

4.1 Требования к Системе в целом

Система должна отвечать следующим требованиям:

1. Обеспечивать возможность ручной загрузки и хранения загружаемых файлов (текстовых документов) с определенным перечнем расширений
2. Производить последовательную обработку (индексацию) успешно загруженных документов с помощью текстового эмбединга с последующей записью обработанных данных в векторную БД.
3. Производить анализ большого объема данных и генерировать ответы на запросы пользователей с использованием большой языковой модели (LLM) в формате естественного общения. Ответы должны быть основаны на данных из ранее загруженных документов с возможностью просмотра фрагмента с оригинальным абзацем из найденного документа.
4. Обеспечивать контекстуальное общение для более детальных запросов.
5. Обеспечивать категоризацию источников по степени их релевантности.
6. Обеспечивать обработку запросов пользователей с получением данных изолированных баз знаний в том числе в режиме оффлайн.
7. Ролевая модель должна быть осуществлена в следующем порядке:
 - Администратор должен иметь полный доступ к управлению модулями системы;
 - Менеджер должен иметь возможность установки правил работы Операторов по базам знаний;
 - Оператор должен иметь возможность ведения интеллектуального диалога с просмотром исходной документации в рамках назначенной ему области знаний.

4.1.1 Структура и функционированию Системы

4.1.1.1 Архитектура системы

Архитектура разработанного ПО удовлетворяет ряду перечисленных ниже требований и была построена в соответствии с дизайном Системы, изображенным на рис.1.

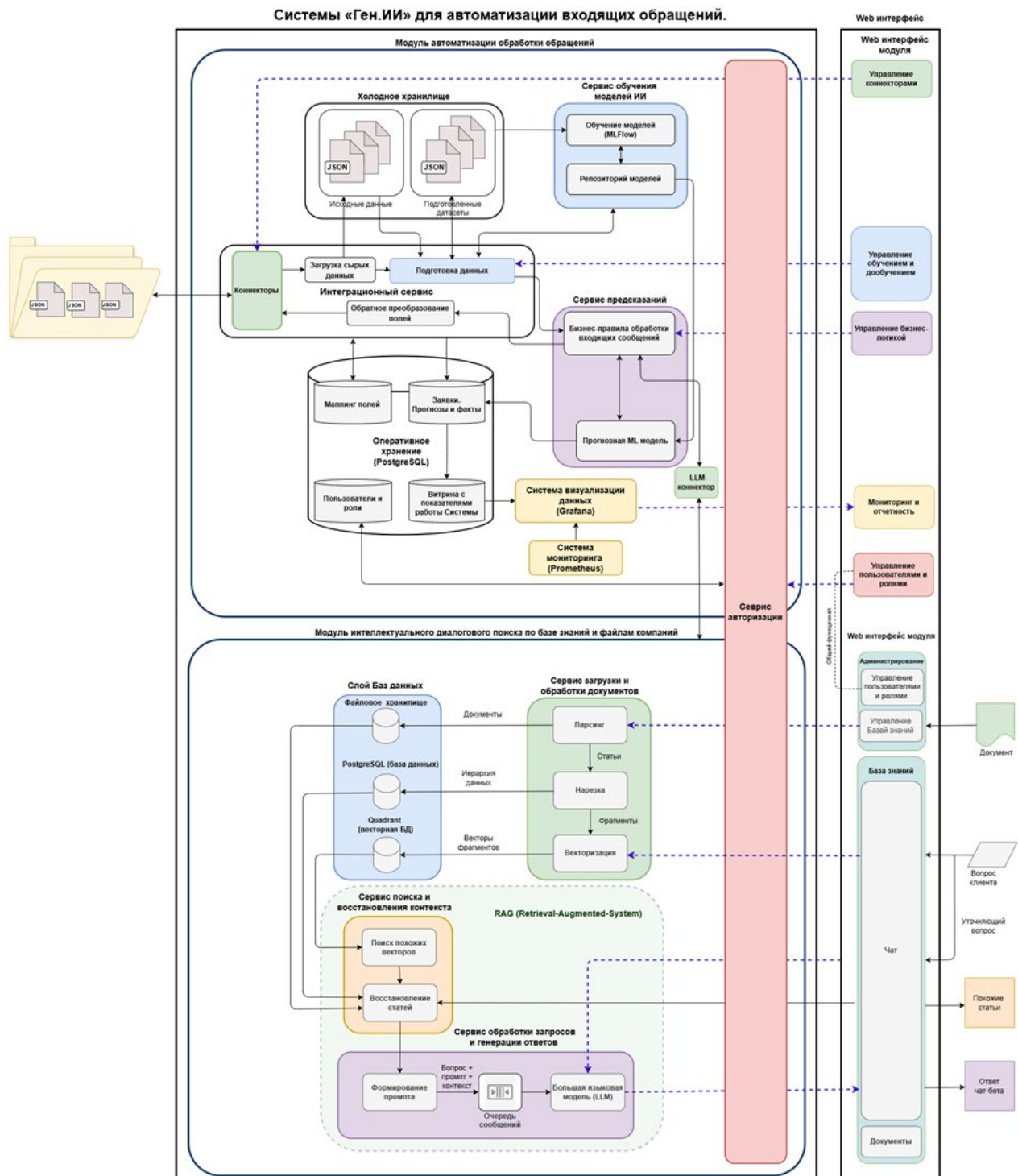


Рис.1 Архитектура ПО

1. При разработке используется только ПО, внесенное в Единый реестр российских программ для электронных вычислительных машин и баз данных, а также свободное

кроссплатформенное ПО (.NET Core, Linux, Postgres, Docker и т.д.);

2. Используются открытые протоколы взаимодействия OpenAPI, HTTP(S), REST, WebSocket;

3. В основу разработки положен сервис-ориентированный подход и микросервисная архитектура;

4. Архитектура системы открыта к расширению за счёт подключения плагинов через API и доработки бизнес-логики через новые микросервисы;

5. Система имеет ролевую модель доступа.

4.1.1.2 Требования к функциональным модулям системы, их назначение и основные характеристики

Разработанное ПО состоит из нескольких основных функциональных модулей, схематично изображенных на функциональной диаграмме на рис.2.

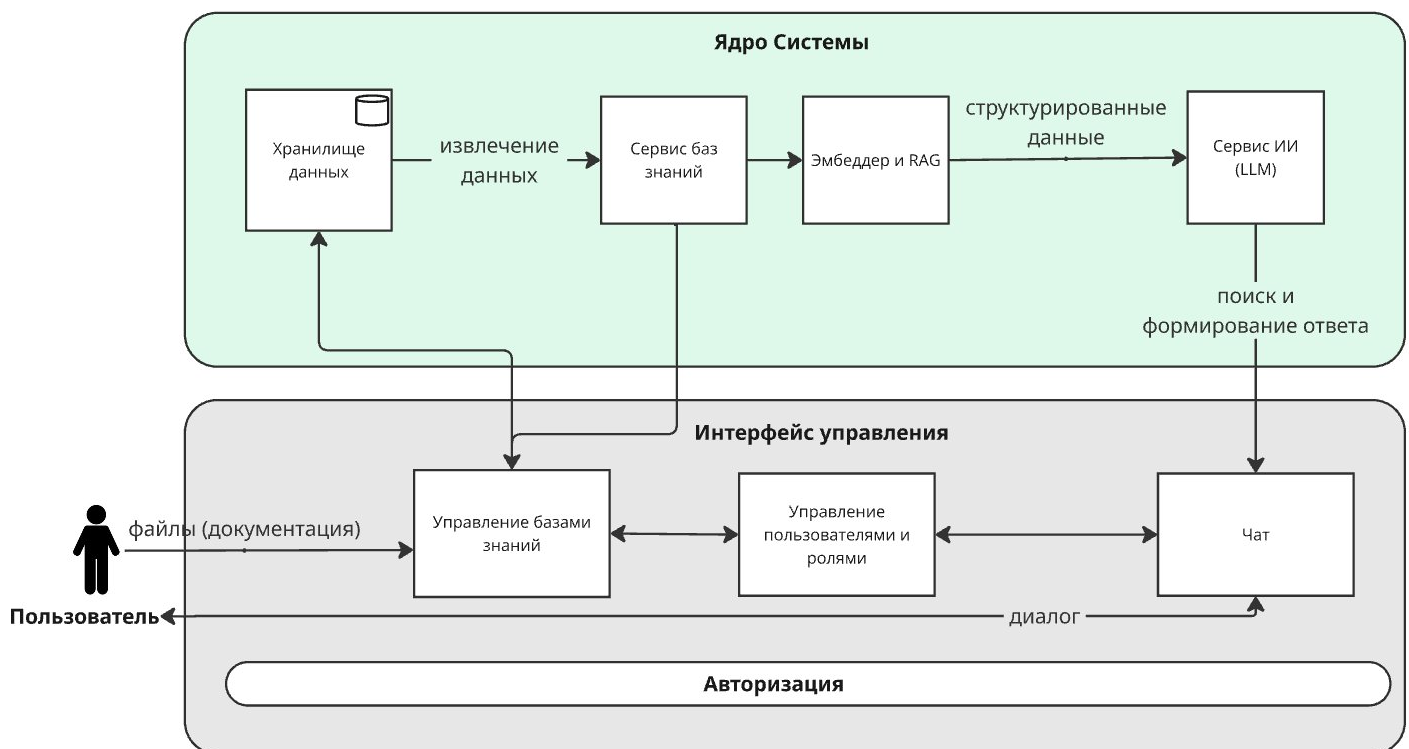


Рис.2. Функциональная диаграмма ПО

Хранилище данных

Функциональный модуль, который отвечает за хранение всех видов файлов, а именно:

- документы, загружаемые пользователями в базы знаний следующих типов расширений: .docx, .pdf, .txt, .pptx. (холодное хранилище);
- документы, разделенные на фрагменты до векторизации;
- векторные представления из извлеченных фрагментов (векторная база данных);
- пользователи и их параметры

Детальное представление архитектуры модуля и его составных частей приведено в разделе 4.1.1.1.

Сервис баз знаний (управление базами знаний)

Функциональный модуль, который отвечает за контент (наполнение) источников данных необходимой информацией для осуществления интеллектуального диалога в чате. Основные задачи модуля:

- обеспечить пользователей возможностью просмотра данных существующих в системе баз знаний и создания новых,

- обеспечить управление документами в каждой из созданных баз знаний, включая множественную загрузку (добавление), удаление отдельных документов и их экспорт.

Детальное представление архитектуры модуля и его составных частей приведено в разделе 4.1.1.1.

Сервис текстового эмбединга и RAG

Функциональный модуль, производящий основное преобразование текста в числовые векторы, которые на последующем этапе помогают модели искусственного интеллекта распознавать смысл и взаимосвязи в тексте для быстрого семантического поиска. Преобразованные слова, предложения или их части от исходного текста переводятся в многомерные векторы, в которых каждое измерение отражает характеристику или связь с другими элементами текста. Учёт контекста при этом помогает модели ИИ точно различать слова с разными значениями.

Основные задачи RAG:

- извлечения релевантной информации из больших объемов данных в базах знаний с помощью использования механизмов поиска,

- генерация ответов на основе извлеченной информации и контекста запроса с использованием языковой модели (LLM),

- объединение результатов извлечения и генерации для контекстуализированных ответов на запросы пользователей.

Детальное представление архитектуры модуля и его составных частей приведено в разделе 4.1.1.1.

Сервис ИИ (LLM), Чат

Ключевой функциональный модуль системы и центральный элемент платформы, обеспечивающий интеллектуальную обработку запросов, генерацию ответов и ведение диалога. Модуль LLM отвечает за создание текстовых ответов на запросы пользователей, используя как факты определенной базы знаний (обученные знания), так и информацию, полученную из модуля RAG, чтобы сформировать точные и контекстуально релевантные ответы.

Детальное представление архитектуры модуля и его составных частей приведено в разделе 4.1.1.1.

Система ролевого доступа (Управление пользователями и ролями)

Система ролевого доступа в платформе отвечает за управление доступом пользователей к различным функциональным возможностям системы. Она обеспечивает безопасность, контроль и управление правами пользователей в зависимости от их ролей. В данной платформе выделяются три основные роли: Оператор, Менеджер и Администратор. Каждая из этих ролей имеет свои функции и задачи.

Детальное описание архитектуры модуля и их составных частей приведены в разделе 4.1.1.1

4.1.1.3 Перспективы развития, модернизации Системы

Программные архитектурные решения, применяемые при разработке Системы, должны обеспечивать возможность дальнейшего развития Системы:

- Доработка блока управления базами знаний и документацией;
- Оптимизации, связанные с генерацией ответов и тонких настроек параметров языковой модели;
- Оценка качества полученных ответов, формирование и визуализация отчетности на основании статистических данных по взаимодействию пользователей с интеллектуальным чатом;
- Доработка ролевой модели (гибкое управление политиками доступа);
- Возможность выбора моделей ИИ и переключения между ними для ведения интеллектуального диалога;
- Настройка дополнительных интеграций для автоматической загрузки документации в базы знаний и доработка модуля интеграций для ведения диалога в мессенджерах.

В случае необходимости модернизации Системы должно быть разработано частное техническое задание на соответствующий блок.

4.1.2 Надежность

Для программного продукта, устанавливаются следующие количественные значения показателей надежности (но не более, чем показатели надежности ЦОД или других аппаратных мощностей, на которых развернута Система):

- режим работы Системы – 24/7 (7 дней в неделю 24 часа в сутки);
- допустимое максимальное время восстановления работоспособности при любых сбоях и отказах не должно превышать 48 (сорока восьми) часов, без учета времени восстановления работоспособности вышедшего из строя оборудования Операционной Системы;
- обслуживание (обновление) Системы должно проводиться в утверждённые окна обслуживания не чаще 1 раза в месяц, за исключением критических обновлений.

4.1.3 Безопасность

Безопасность текущей инфраструктуры обеспечивается за счет использования следующих элементов инфраструктуры:

- Role-Based Access Control (RBAC) позволяет разделять права пользователей и делегировать административные задачи, поддерживая высокие стандарты безопасности;
- Public Key Infrastructure (PKI) использует сертификаты, выданные доверенными центрами сертификации (ЦС) для аутентификации серверов и обеспечения целостности данных;
- Transport Layer Security (TLS), HTTPS over SSL (HTTPS);
- Стандартных протоколов для аутентификации пользователей, где это возможно.

Эти ключевые элементы безопасности применяются для идентификации доверенных пользователей, серверов, подключений и операций, что обеспечивает надежную основу для инфраструктуры.

4.1.4 Эргономика и техническая эстетика

4.1.4.1 Общие требования к эргономике и технической эстетике

В графическом веб-интерфейсе ввод-вывод данных системы, прием управляющих команд и отображение результатов их исполнения выполняются в интерактивном режиме. В случае

длительных процессов обработки информации (например, экспорта или загрузки документов в базы знаний) предусмотрена индикация хода процесса. Платформа не поддерживает адаптивный дизайн для различных типов экранов и расширений.

4.1.4.2 Требования к эргономике и технической эстетике пользовательского интерфейса решения

Экранные формы программных компонент спроектированы с учетом требований унификации, обеспечивая положительный пользовательский опыт:

- ✓ Интуитивно понятный интерфейс: все экранные формы выполнены в едином графическом дизайне (в пределах контура Системы), с одинаковым и логичным расположением основных элементов управления и навигации;
- ✓ для обозначения сходных операций (в пределах контура Системы) используются исходные графические элементы (знакомые и понятные иконки и обозначения). Термины, используемые для обозначения типовых операций, а также последовательности действий при их выполнении, унифицированы;
- ✓ Читаемость и доступность: применение шрифтов, обеспечивающих хорошую читаемость с контрастностью текста и фона для удобства восприятия информации;
- ✓ Эстетика и визуальная гармония: согласованная цветовая палитра, соответствующая тематике платформы;
- ✓ использование списков для удобного отображения информации;
- ✓ Тестирование и улучшение: регулярное тестирование интерфейса для выявления проблем и улучшения эргономики;
- ✓ сбор обратной связи для постоянного совершенствования пользовательского опыта
- ✓ Наличие доступной документации и справочных материалов для пользователей

Экранные формы корректно отображаются при разрешении экрана 1280x1024 пикселей и выше. Интерфейс Системы реализован на русском языке.

4.1.4.3 Защита информации от несанкционированного доступа

Удаленный доступ к ресурсам Системы осуществляется только после аутентификации и авторизации пользователя.

Средства авторизации и аутентификации обеспечивают проверку введенных идентификационных данных, роль пользователя в системе определяется в соответствии с данными из заведенного в Системе каталога учетных записей пользователей.

4.2 Описание функциональных блоков Системы

Система включает в себя несколько ключевых функциональных блоков: «Чат», «Управление базами знаний» и «Управление пользователями», доступ к которым осуществляется через боковое меню интерфейса.

Работа всех разделов системы осуществляется в контексте активной базы знаний. Ее выбор возможен автоматически при входе пользователя или вручную - через выпадающий список в верхней части интерфейса. В зависимости от выбранной базы, изменяется содержимое разделов «Чат» и «Документы». Ниже приведено описание основных функциональных блоков системы.

4.2.1 Описание функционального блока «Модуль управления базами знаний»

В данном разделе осуществляются функции, которые обеспечивают эффективное управление, организацию и доступ к данным, используемым для интеллектуального поиска и последующей генерации ответов:

- создание и удаление баз знаний;
- добавление (путем множественной загрузки файлов) и удаление документов из базы знаний;
- экспорт необходимых документов из базы знаний;
- обеспечение взаимодействия с холодным хранилищем для хранения больших объемов данных;
- обеспечение безопасности данных и конфиденциальности информации, хранящейся в базах знаний.

Рабочая зона - центральное место занимает таблица со списком баз знаний. Для каждой базы отображаются:

- Название
- Количество документов

В правой части каждой строки - иконка удаления Базы знаний (с подтверждением действия).

В верхней части располагается кнопка «Создать», которая открывает форму создания базы знаний с двумя обязательными для заполнения полями:

- Название
- Описание

После создания базы знаний Пользователю предоставляется возможность загрузить документацию. Для этого необходимо перейти в выбранную Базу знаний и нажать на «Документы» в боковом меню.

Верхняя часть страницы с выбранной базой содержит название активной базы знаний и кнопку «Загрузить». После нажатия открывается окно выбора файлов. Система поддерживает:

- Загрузку до 100 файлов одновременно
- Размер каждого файла не более 50 МБ
- Форматы: PDF, DOCX, TXT

После подтверждения загрузки документы появляются в таблице, где отображаются:

- Название
- Формат

Рядом с каждым документом находятся кнопки:

- Скачать
- Удалить (требует подтверждения)

Загрузка и удаление документов доступны администраторам и менеджерам.

Чтобы начать диалог на основе созданной базы знаний с загруженными документами Пользователю необходимо вернуться к списку Баз знаний, нажав на боковом меню «Базы знаний».

При выборе определенной базы знаний из списка Пользователь переходит в модуль «Чат».

4.2.2 Описание функционального блока «Чат»

Основной раздел взаимодействия пользователя с выбранной базой знаний, позволяющий формулировать вопросы и получать ответы, сформированные на основе загруженного контента. Интерфейс построен в формате диалога и обеспечивает интуитивно понятную работу с системой. Пользователю предоставляется следующий функционал:

- ввод вопроса в текстовое поле и отправка запроса;
- отображение связанного текстового ответа, сформированного на основе документов;
- просмотр использованных источников с указанием названия документа, текста фрагмента и степени релевантности в процентах (включая исходные изображения, содержащиеся в документе);
- возможность скачать исходный документ по клику на название в списке источников;
- поддержка многоходового диалога с учетом контекста предыдущих сообщений.

Рабочая зона чата разделена на две области:

1. Левая часть - зона диалога с полем ввода текста вопроса. Кнопка «Задать вопрос» активируется при вводе текста.

Кнопки:

- «Продолжить общение» - сохраняет диалог и позволяет задать следующий вопрос в рамках текущего контекста.

- «Очистить» - удаляет историю и начинает новый диалог. Доступна в любой момент.

2. Правая часть - фрагменты документов

- Отображаются фрагменты текста, на основе которых был сформирован ответ
- Отображается степень релевантности в процентах
- Указывается название документа
- При клике на ссылку с названием документа инициируется экспорт исходного документа.

4.2.3 Описание функционального блока «Управление пользователями»

Данный функциональный блок предназначен для администрирования учетных записей пользователей и настройки их ролей. В интерфейсе отображается таблица всех зарегистрированных пользователей с основными данными. Пользователю доступны следующие действия:

- создание новой учетной записи;
- редактирование информации о пользователе и назначенных ролях;
- удаление учетной записи с подтверждением действия;
- автоматическое ограничение доступа на уровне интерфейса в зависимости от назначенной роли.

Рабочая зона: центральная таблица отображает список всех пользователей с данными:

- Имя
- Email
- Назначенные роли

В правой части каждой строки - иконка редактирования и удаления пользователя (с подтверждением действия).

В верхней части - кнопка «Создать», открывающая форму добавления пользователя, которая содержит следующие данные для заполнения:

- Имя
- Email
- Пароль (с возможностью автогенерации уникального пароля)
- Назначение ролей (администратор, менеджер, оператор)

Данный раздел доступен только администраторам.

4.3 Техническое обеспечение

Для соблюдения требований по надежности, эксплуатации и техническому обслуживанию Системы, должны быть предоставлены аппаратно-программные средства, необходимые для настройки и функционирования Системы.

Для соблюдения требований по надежности, эксплуатации и техническому обслуживанию Системы, были предоставлены аппаратно-программные средства, необходимые для настройки и функционирования Системы.

Рабочая среда серверной части Системы включает следующие программные средства:

- PostgreSQL - СУБД для реализации системы хранения информационного наполнения Системы;
- Quadrant – векторная СУБД для хранения больших объемов векторных данных;
- Astra Linux - операционная система сервера, на котором будет развернута Система;
- сетевой канал с пропускной способностью не менее 10 Мбит/сек.

Для использования и администрирования Системы компьютер пользователя должен соответствовать следующим требованиям:

- операционная система Windows 7 или старше или Linux с графическим интерфейсом;
- оперативная память компьютера не менее 2 Гб;
- наличие установленного интернет-обозревателя (браузера):
- Edge версии не ниже 102;
- FireFox версии не ниже 100;
- Яндекс Браузер 22.0+;
- Google Chrome версии не ниже 102

В браузере разрешено использование javascript.

4.4 Входные и выходные данные

4.4.1 Входные данные

Входными данными являются загружаемые пользователем документы – источники данных для интеллектуального диалога. Документы должны быть определенного формата: PDF, DOCX, TXT, при этом максимальный размер одного файла не должен превышать 50 Мб. За одну операцию можно загрузить не более 100 файлов.

4.4.2 Выходные данные

Выходными данными являются автоматически сгенерированные ответы на пользовательские запросы. Они представляют собой текст на естественном языке, сформированный на основе обработанных документов базы знаний. Ответ сопровождается списком использованных фрагментов документов с указанием названия документа, текста фрагмента и степени релевантности в процентах.