Описание процедуры эксплуатации Программного обеспечения «Системы интеллектуального поиска по базе знаний с использованием технологии Retrieval-Augmented Generation (RAG)»

2026 г.

Содержание

1. Назначение системы

Программное обеспечение представляет собой корпоративную информационную систему, реализующую интеллектуальный поиск по базе знаний организации с использованием технологии Retrieval-Augmented Generation (RAG).

Система предназначена для обработки текстовых запросов пользователей на естественном (русском) языке и формирования ответов на основе проиндексированных данных корпоративной Wiki-базы знаний.

Основное назначение системы — обеспечение оперативного доступа сотрудников к регламентам, инструкциям, методическим материалам и иным внутренним документам организации посредством единого веб-интерфейса.

2. Общие положения

2.1. Область применения

Требования настоящей процедуры эксплуатации распространяются на программное обеспечение, используемое во внутренней информационной инфраструктуре организации.

Система применяется для обеспечения сотрудников оперативным доступом к информации, содержащейся в корпоративной базе знаний, в рамках их служебной деятельности.

Пользователями системы являются:

- Сотрудники организации, осуществляющие поиск информации в базе знаний;
- Администраторы системы, осуществляющие настройку и сопровождение программного обеспечения;
- Специалисты ИТ-подразделения, обеспечивающие техническую поддержку, сопровождение и модернизацию системы.

Доступ к системе осуществляется через веб-интерфейс в пределах корпоративной сети либо по защищённым каналам связи в соответствии с установленной политикой информационной безопасности организации.

2.2. Термины, определения и сокращения

API (Application Programming Interface) — программный интерфейс взаимодействия между компонентами системы, обеспечивающий передачу данных и выполнение функциональных операций.

Backend — серверная часть программного обеспечения, обеспечивающая обработку пользовательских запросов, взаимодействие с базами данных и интеграцию с языковой моделью.

Frontend — пользовательский веб-интерфейс системы, обеспечивающий ввод запросов и отображение результатов обработки.

LLM (Large Language Model) — большая языковая модель, предназначенная для обработки и генерации текстовой информации на естественном (русском) языке.

RAG (Retrieval-Augmented Generation) — архитектурный подход, при котором генерация ответа языковой моделью осуществляется с использованием предварительно найденных релевантных фрагментов из базы знаний.

База знаний (Wiki) — корпоративный источник структурированной и неструктурированной информации (регламенты, инструкции, методические материалы), используемый системой в качестве источника данных для поиска.

Векторное представление (эмбеддинг) — числовое представление текстового

фрагмента, используемое для определения семантической близости между запросом пользователя и данными базы знаний.

Векторное хранилище — специализированное хранилище данных, предназначенное для хранения векторных представлений текстовых фрагментов и выполнения семантического поиска.

Индексация — процесс предварительной обработки данных базы знаний, включающий разбиение текстов на фрагменты и подготовку их для последующего поиска.

Контейнеризация — способ развертывания программного обеспечения с использованием изолированных контейнеров, содержащих приложение и его зависимости.

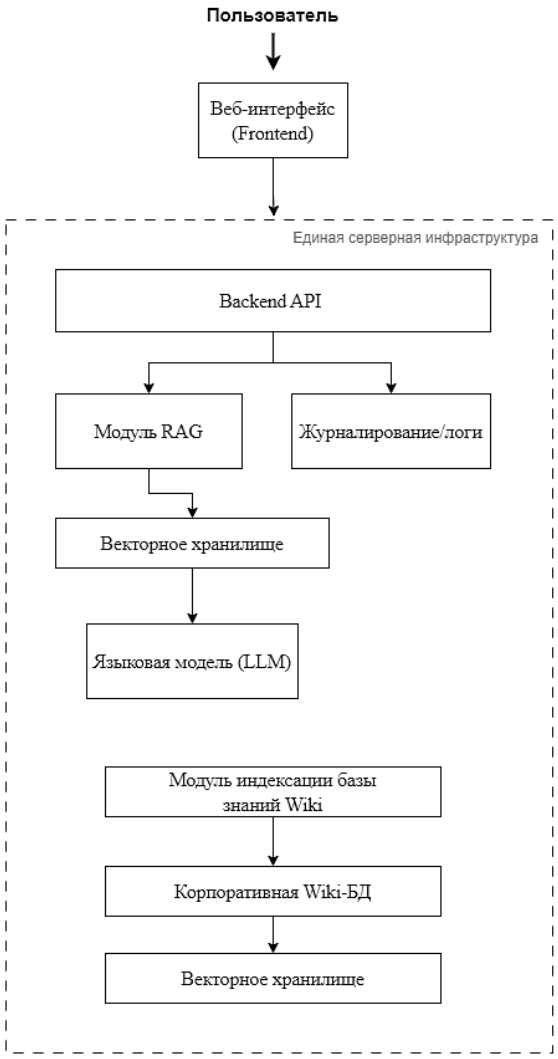
Семантический поиск — поиск информации на основе смысловой близости текстов, а не точного совпадения слов.

3. Процедура эксплуатации

3.1 Структура системы

Структурная схема программного обеспечения представлена на рисунке 1.

Рисунок 1 – Структурная схема системы



1. Пользовательский веб-интерфейс (Frontend) — обеспечивает ввод текстовых запросов пользователем и отображение сформированных системой ответов.

2. Серверная часть (Backend API) — осуществляет приём и обработку запросов от веб-интерфейса, координирует взаимодействие между модулем RAG, векторным хранилищем, LLM-сервисом и вспомогательными компонентами системы.

3. Модуль RAG — реализует логику семантического поиска релевантных фрагментов в векторном хранилище и формирует расширенный запрос к LLM-сервису с учётом найденного контекста.

4. LLM-сервис (языковая модель) — развернут как отдельный компонент серверной инфраструктуры и выполняет генерацию текстового ответа на основе пользовательского запроса и переданного контекста.

5. Модуль индексации базы знаний — обеспечивает получение данных из корпоративной

Wiki, их предварительную обработку, формирование векторных представлений и обновление векторного индекса.

6. Векторное хранилище — хранит векторные представления текстовых фрагментов базы знаний и используется модулем RAG для выполнения семантического поиска.

7. Подсистема журналирования и служебная база данных — обеспечивает хранение служебной информации, логирование операций и поддержку мониторинга функционирования системы.

8. Взаимодействие компонентов осуществляется через внутренние программные интерфейсы в пределах единой серверной инфраструктуры организации; модуль индексации взаимодействует с Wiki и векторным хранилищем, а LLM-сервис используется исключительно на этапе формирования ответа пользователю.

3.2. Эксплуатация системы

3.2.1. Описание ролей и прав

В системе реализована ролевая модель разграничения доступа. Права пользователей определяются их ролью.

Описание ролей

Пользователь — сотрудник организации, использующий систему для выполнения поисковых запросов и получения ответов.

Администратор — сотрудник, осуществляющий настройку индексации, управление доступом пользователей и контроль функционирования системы на прикладном уровне.

Технический администратор — специалист ИТ-подразделения, обеспечивающий развертывание, обновление и мониторинг серверной инфраструктуры системы.

3.2.2. Начало работы пользователя с системой

Далее будут описаны целевые сценарии использования системы ролью «Пользователь». Как пользователь системы, я могу начать работу в системе для выполнения поиска по базе знаний.

Предусловия:

1. Пользователь имеет действующую учетную запись в корпоративной системе.
2. Пользователь имеет права доступа к программному обеспечению.

Список действий:

1. Открыть веб-интерфейс системы в браузере.
2. При необходимости пройти процедуру авторизации в соответствии с корпоративной политикой доступа.
3. После успешной авторизации перейти на главный экран системы.
4. Убедиться в доступности поля ввода запроса.

Результат:

1. Пользователь получает доступ к функционалу системы в соответствии со своей ролью.
2. В интерфейсе отображается активное поле для ввода запроса.
3. Система готова к приёму пользовательских запросов.

3.2.3. Выполнение поиска по базе знаний

Как пользователь системы, я могу получить ответ на вопрос по внутренним регламентам и материалам компании.

Предусловия:

1. Пользователь имеет действующую учетную запись в корпоративной системе.
2. Пользователь авторизован в системе.
3. Система функционирует в штатном режиме.

Список действий:

1. Открыть веб-интерфейс системы в браузере.
2. Ввести текстовый запрос в поле ввода (например, сформулировать вопрос по регламенту или инструкции).
3. Нажать кнопку отправки запроса либо клавишу Enter.
4. Дождаться обработки запроса системой.
5. Ознакомиться с полученным ответом, отображённым в интерфейсе.

Результат:

1. Система выполняет семантический поиск релевантной информации в базе знаний.
2. На основе найденных данных формируется текстовый ответ.
3. Ответ отображается пользователю в диалоговом интерфейсе.

4. При отсутствии релевантной информации система уведомляет пользователя о невозможности сформировать корректный ответ.

3.2.4. Выполнение повторного (уточняющего) запроса

Как пользователь системы, я могу уточнить ранее заданный вопрос для получения более точного ответа.

Предусловия:

1. Пользователь авторизован в системе.
2. Ранее был выполнен первичный запрос и получен ответ системы.
3. Диалоговая сессия пользователя активна.

Список действий:

1. Ознакомиться с ранее полученным ответом системы.
2. Ввести уточняющий или дополнительный вопрос в поле ввода сообщения.
3. Нажать кнопку отправки запроса либо клавишу Enter.
4. Дождаться обработки запроса системой.
5. Ознакомиться с обновлённым ответом, отображённым в интерфейсе.

Результат:

1. Система учитывает контекст предыдущего запроса.
2. Выполняется дополнительный семантический поиск релевантной информации.
3. Формируется уточнённый ответ с учётом нового запроса и контекста диалога.
4. Ответ отображается пользователю в диалоговом интерфейсе.

3.2.5. Завершение работы пользователя в системе

Как пользователь системы, я могу завершить работу в системе после выполнения необходимых действий.

Предусловия:

1. Пользователь авторизован в системе.

Список действий:

1. Завершить текущую сессию работы с системой.
2. При наличии соответствующей функции — нажать кнопку «Выход» (Logout).
3. Закрыть окно браузера либо вкладку с веб-интерфейсом системы.

Результат:

1. Пользовательская сессия завершается.
2. Доступ к системе прекращается до повторной авторизации.
3. Данные пользовательской сессии сохраняются или очищаются в соответствии с политикой информационной безопасности организации.