Документация Jay Knowledge Hub

Начало работы с Jay Knowledge Hub

Jay Knowledge Hub от Just AI — это мощный инструмент для управления знаниями, который поможет вам эффективно их организовать и использовать. Это не просто хранилище информации. С помощью Jay Knowledge Hub вы сможете создавать умных чат-ботов, автоматизировать ответы на часто задаваемые вопросы и значительно повысить уровень обслуживания ваших клиентов и сотрудников.

Подробнее о продукте Jay Knowledge Hub можно узнать на нашем сайте.

Шаг 1. Регистрация

Зарегистрируйтесь в Jay Knowledge Hub. Если вы уже работали с другими продуктами Just AI, вы можете использовать свою прежнюю учетную запись для входа на платформу.

Шаг 2. Создание проекта и загрузка данных

Создайте ваш первый проект и загрузите данные в Jay Knowledge Hub. Используйте файлы форматов: PDF, DOCX, TXT, XML, JSON, YAML, MD, MDX, HTML, CSV.

Если ваши знания хранятся в Confluence, Minerva Knowledge или на веб-сайте, вы можете настроить интеграцию с Jay Knowledge Hub и загружать данные автоматически.

Шаг 3. Индексация базы знаний

Запустите индексацию базы знаний на загруженных данных. Процесс индексации может занять некоторое время — следите за прогрессом в интерфейсе продукта. Затем вы сможете использовать Jay Knowledge Hub для поиска информации по вашим данным.

Шаг 4. Тестирование

Перейдите в раздел *Тестирование*, чтобы проверить работу базы знаний. Попробуйте разные запросы, чтобы убедиться, что Jay Knowledge Hub работает корректно. В тестовом диалоге вы можете экспериментировать с настройками поиска и подготовки ответов. Если результат устраивает, настройки можно сохранить в проект прямо из диалога.

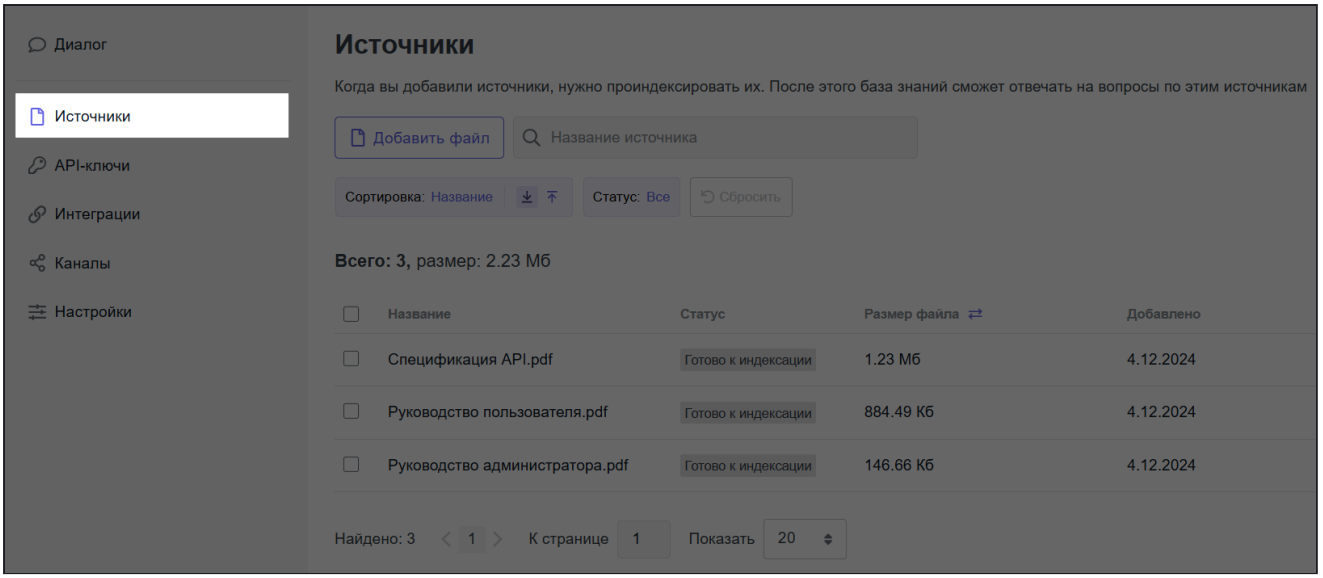
Шаг 5. Доступ к вашим знаниям

Чтобы ваши клиенты и сотрудники могли задавать вопросы к базе знаний:

- Подключите каналы коммуникации: чат-виджеты, мессенджеры и другие интерфейсы.
- Интегрируйте базу знаний с системой CRM, Helpdesk и другими инструментами с помощью API.
- Если вы используете AI-ассистент Jay Copilot, подключите базу знаний в качестве приложения.

Загрузка файлов вручную

Чтобы загрузить данные в виде файлов, перейдите в раздел *Источники*. Если при создании проекта вы выбрали *Загрузить файлы вручную*, то попадете в этот раздел автоматически.



Требования к файлам

Сейчас в Jay Knowledge Hub можно загружать файлы, если они соответствуют требованиям:

- Допустимые форматы: PDF, DOCX, TXT, XML, JSON, YAML, MD, MDX, HTML, CSV.
- Максимальное количество: 100 штук.
- Максимальное количество файлов CSV: 10 штук.
- Размер одного файла: до 25 Мб.

<details>
<summary>Требования к формату CSV</summary>

- Первая строка должна содержать названия столбцов.
- Допустимые разделители: `,` или `;`.
</details>

💡 Рекомендации по загрузке

- Проверьте, что файлы не содержат конфиденциальную информацию. Все загруженные данные будут доступны для индексации и использования. Ваши клиенты или сотрудники

смогут получить к ним доступ, когда будут обращаться к базе знаний.

- Используйте *текстовые* файлы, где текст понятен системе. Например, сканированные документы не всегда хорошо распознаются. Это может повлиять на процесс индексации и качество ответов пользователям вашей базы знаний.
- В окне загрузки после выбора файлов появляется возможность указать сегменты базы знаний, к которым будут относиться загружаемые файлы. Сегменты позволяют сгруппировать источники с общей темой. В дальнейшем сегменты можно указывать в запросе к базе знаний для поиска информации только по этим сегментам. Подробнее читайте в разделе Сегменты.

К СВЕДЕНИЮ

- В дополнение к файлам вы можете настроить интеграции для загрузки данных из внешних систем.
- После загрузки данных нужно проиндексировать базу знаний.

Интеграция с Confluence

Интеграция позволяет загружать данные из Confluence в базу знаний, автоматически синхронизировать данные и переиндексировать их.

> ⓘ **ПРИМЕЧАНИЕ**
>
> Если Confluence размещается в инфраструктуре вашей компании, необходимо обеспечить сетевой доступ к API Confluence из Jay Knowledge Hub.

Какие данные будут загружены

В базу знаний будут загружены все страницы выбранного пространства Confluence, исключая комментарии. Возможность загрузить только отдельные страницы или исключить отдельные страницы не предусмотрена.

Вместе со страницами будут загружены прикрепленные к ним изображения.

В настройках интеграции можно включить загрузку и других прикрепленных файлов в допустимых форматах. Допустимы те же форматы, что и при загрузке вручную.

Изображения загружаются в любом случае.

> ⚠ **ПРЕДУПРЕЖДЕНИЕ**
>
> Проверьте, что страницы выбранного пространства не содержат конфиденциальную информацию. Все загруженные данные будут доступны для индексации и использования. Ваши клиенты или сотрудники смогут получить к ним доступ, когда будут обращаться к базе знаний.

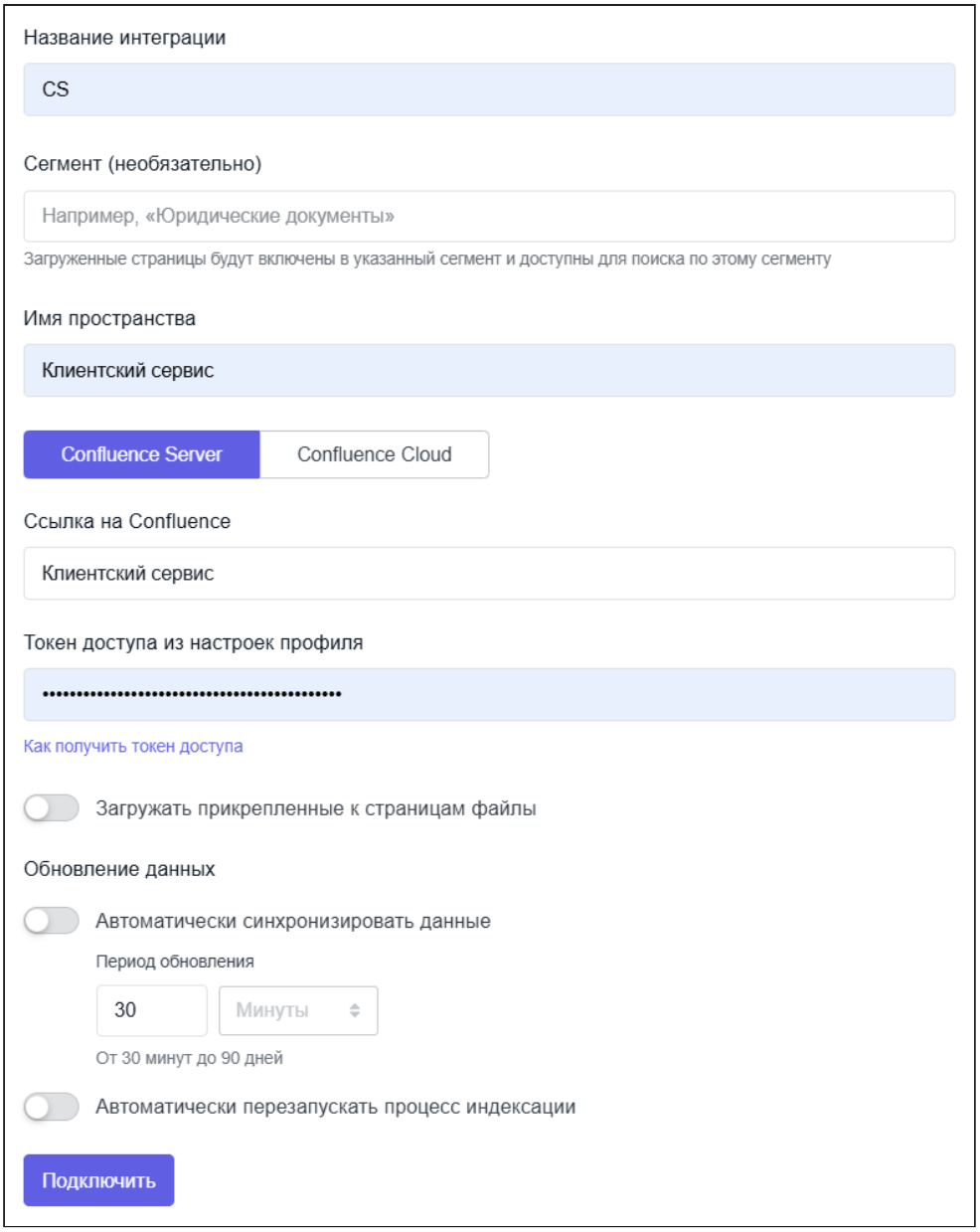
Создание интеграции

Чтобы загрузить данные из Confluence:

1. Перейдите в раздел *Интеграции* и нажмите *Подключить* → *Confluence*.

   Если при создании проекта вы выбрали *Подключить интеграцию с Confluence*, то попадете на страницу новой интеграции автоматически.

2. Укажите настройки интеграции:



- *Название интеграции* — используется как префикс к названиям источников, загруженных из этой интеграции.
- *Сегменты* — к ним будут относиться источники, загруженные из этой интеграции. Сегменты позволяют сгруппировать источники с общей темой. В дальнейшем сегменты можно указывать в запросе к базе знаний для поиска информации только в этим сегментах. Подробнее читайте в разделе Сегменты.
- *Имя пространства* — название пространства или раздела (но не ключ и не идентификатор).
- Параметры для доступа к Confluence:
  - Если вы используете Confluence Data Center или Confluence Server, вставьте ссылку на Confluence, а также токен (маркер) доступа.

    ▼ Как получить токен доступа

a. В Confluence нажмите на свой аватар в правом верхнем углу и перейдите в *Настройки*.

b. В боковой панели выберите *Личные маркеры доступа*.

c. Нажмите *Создать маркер*. Заполните параметры и нажмите *Создать*.

d. Скопируйте токен.

Подробнее о токенах доступа читайте в документации Confluence.

- Если вы используете Confluence Cloud, укажите свой сайт (поддомен) на `atlassian.net`.

- *Загружать прикрепленные к страницам файлы* — включите, чтобы загружать прикрепленные к страницам файлы в допустимых форматах. Эта настройка не влияет на загрузку изображений: они загружаются в любом случае.
- *Автоматически синхронизировать данные* — включите для поддержания актуальности базы знаний, отключите для однократной загрузки.

  > **ПОДСКАЗКА**
  >
  > Не допускайте размещения некорректных или конфиденциальных данных в выбранном пространстве Confluence, ведь они автоматически появятся в базе знаний при очередной синхронизации.

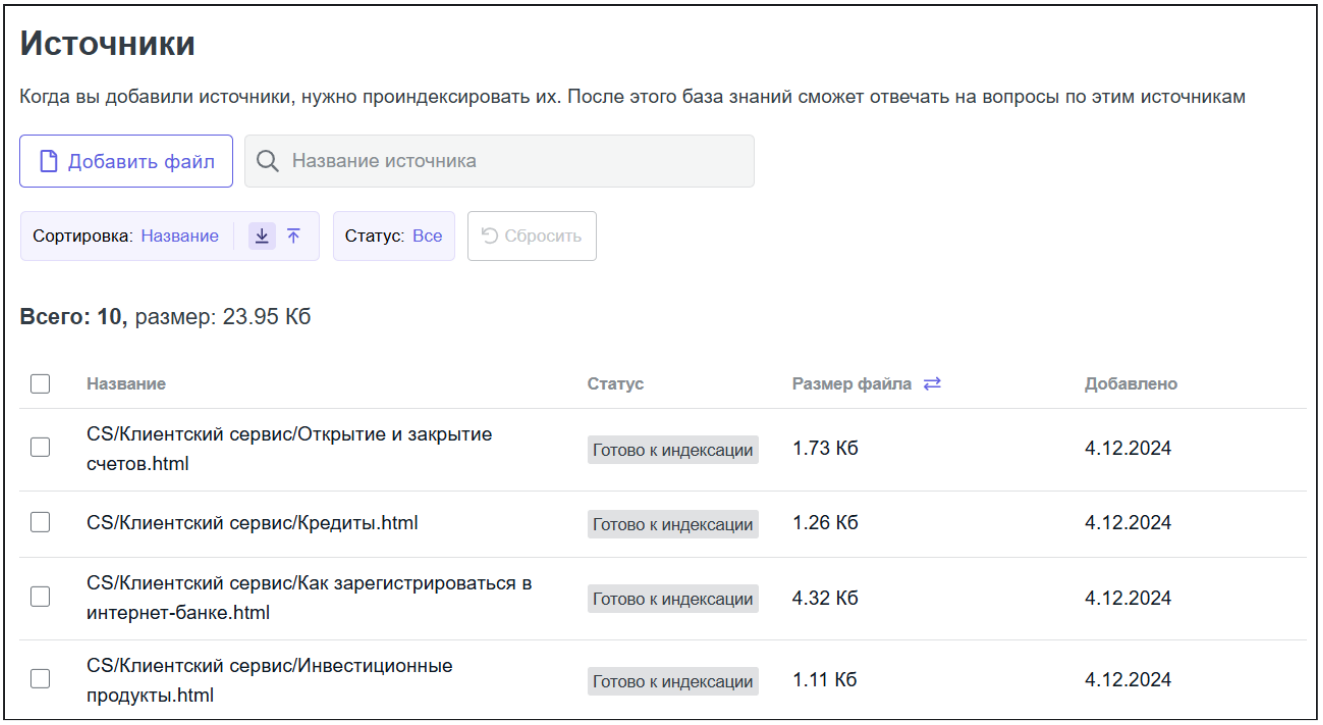
- *Автоматически перезапускать процесс индексации* — включите, чтобы переиндексировать базу знаний после очередной синхронизации.

> **ПРИМЕЧАНИЕ**
>
> Все настройки впоследствии можно изменить, кроме имени пространства и ссылки на Confluence для Confluence Cloud — после подключения они станут недоступны для редактирования.

3. Нажмите *Подключить*.
4. Если вы подключаете Confluence Cloud, в новой вкладке откроется страница запроса доступа. Войдите в Confluence Cloud и разрешите приложению Jay Knowledge Hub доступ к своим данным.

   > **К СВЕДЕНИЮ**
   >
   > Доступ действует в течение года. Затем его нужно будет подтвердить повторно.

Статус интеграции сразу после создания — «Идет обновление». Когда статус сменится на «Подключено», вы можете просмотреть загруженные страницы в разделе *Источники*.



В дополнение к данным из интеграций вы можете загрузить файлы вручную.

После загрузки данных нужно проиндексировать базу знаний.

> **К СВЕДЕНИЮ**
>
> О принудительной синхронизации и изменении настроек читайте в разделе Управление интеграциями.

Интеграция с Minerva Knowledge

Интеграция позволяет загружать данные из Minerva Knowledge в базу знаний, автоматически синхронизировать данные и переиндексировать их.

> ⓘ **ПРИМЕЧАНИЕ**
>
> Если Minerva Knowledge размещается в инфраструктуре вашей компании, необходимо обеспечить сетевой доступ к API Minerva Knowledge из Jay Knowledge Hub.

Какие данные будут загружены

В базу знаний будут загружены все статьи выбранного проекта Minerva Knowledge. Возможность загрузить только отдельные статьи или исключить отдельные статьи не предусмотрена. Контент других типов (шаблоны, документы) игнорируется.

Вместе со статьями будут загружены прикрепленные к ним изображения. Дополнительно вы можете включить в настройках интеграции загрузку других прикрепленных файлов в допустимых форматах. Допустимы те же форматы, что и при загрузке вручную.

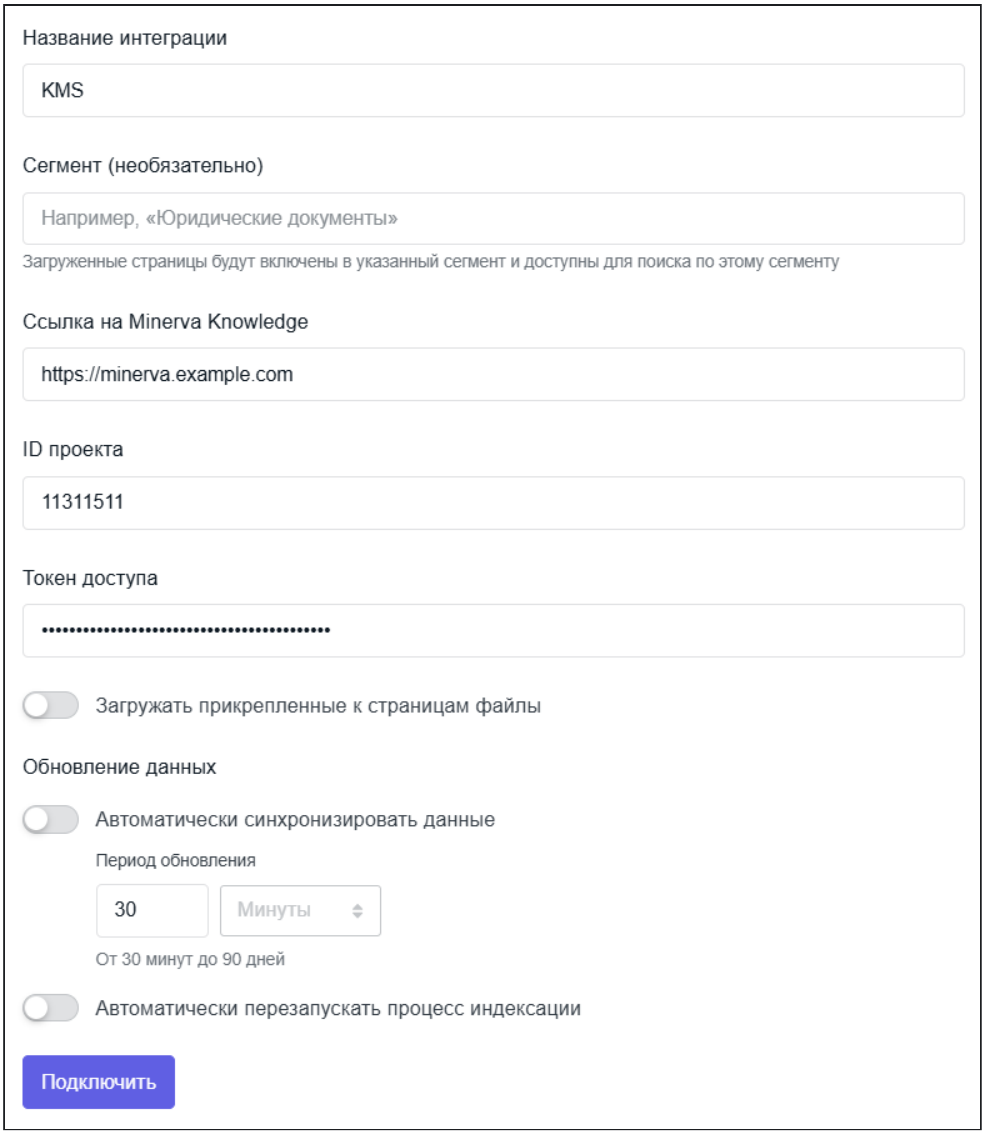
> ⚠ **ПРЕДУПРЕЖДЕНИЕ**
>
> Проверьте, что статьи выбранного проекта не содержат конфиденциальную информацию. Все загруженные данные будут доступны для индексации и использования. Ваши клиенты или сотрудники смогут получить к ним доступ, когда будут обращаться к базе знаний.

Создание интеграции

Чтобы загрузить данные из Minerva Knowledge:

1. Перейдите в раздел *Интеграции* и нажмите *Подключить → Minerva Knowledge*.

   Если при создании проекта вы выбрали *Подключить интеграцию*, то попадете на страницу новой интеграции автоматически.
2. Укажите настройки интеграции:



- *Название интеграции* — используется как префикс к названиям источников, загруженных из этой интеграции.
- *Сегменты* — к ним будут относиться источники, загруженные из этой интеграции. Сегменты позволяют сгруппировать источники с общей темой. В дальнейшем сегменты можно указывать в запросе к базе знаний для поиска информации только в этим сегментах. Подробнее читайте в разделе Сегменты.
- *Ссылка на Minerva Knowledge* — вставьте домен, где размещена Minerva Knowledge.
- *ID проекта* — например, если адрес проекта — `https://{yourDomain}/content/space/{projectId}`, то `{projectId}` — его идентификатор.
- *Токен доступа* — как сгенерировать токен, читайте в руководстве системного администратора Minerva Knowledge.
- *Загружать прикрепленные к страницам файлы* — включите, чтобы загружать прикрепленные к страницам файлы в допустимых форматах. Эта настройка не влияет на загрузку изображений: они загружаются в любом случае.
- *Автоматически синхронизировать данные* — включите для поддержания актуальности базы знаний, отключите для однократной загрузки.

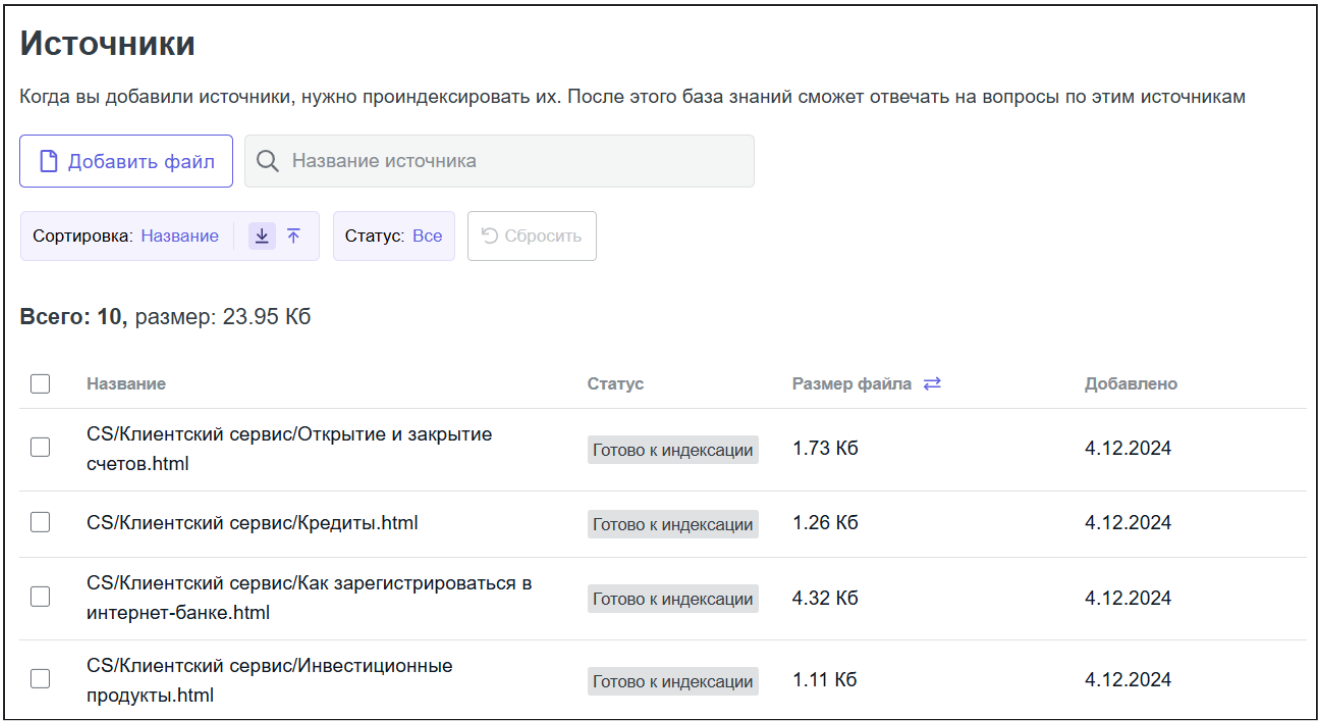
> **ПОДСКАЗКА**

> Не допускайте размещения некорректных или конфиденциальных данных в выбранном проекте, ведь они автоматически появятся в базе знаний при очередной синхронизации.

- *Автоматически перезапускать процесс индексации* — включите, чтобы переиндексировать базу знаний после очередной синхронизации.

> **ⓘ ПРИМЕЧАНИЕ**
>
> Все настройки впоследствии можно изменить.

Статус интеграции сразу после создания — «Идет обновление». Когда статус сменится на «Подключено», вы можете просмотреть загруженные страницы в разделе *Источники*.



В дополнение к данным из интеграций вы можете загрузить файлы вручную.

После загрузки данных нужно проиндексировать базу знаний.

> **ⓘ К СВЕДЕНИЮ**
>
> О принудительной синхронизации и изменении настроек читайте в разделе Управление интеграциями.

Интеграция с веб-сайтом

Интеграция позволяет загружать данные с веб-сайта в базу знаний, автоматически синхронизировать и переиндексировать их.

Обработка страниц из интеграции с веб-сайтом тарифицируется отдельно.

Какие данные будут загружены

Предусмотрено два варианта загрузки:

- Указанные вами страницы.
- Страницы, найденные роботом при обходе сайта, начиная с главной страницы.

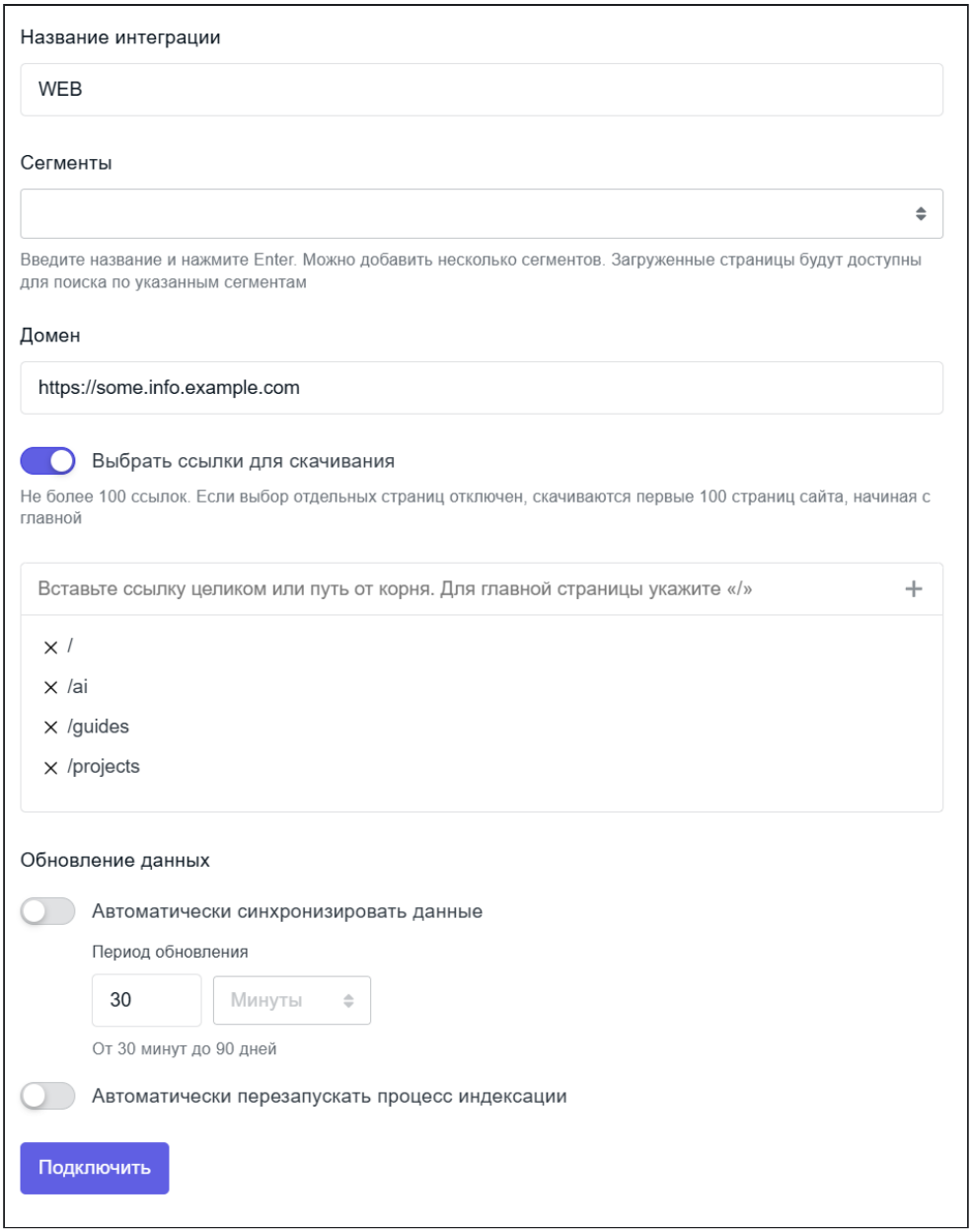
> ⚠ **ОГРАНИЧЕНИЯ**
>
> - Общее количество страниц — не более 100.
> - Загружается только текст HTML-страниц. Изображения не загружаются и не обрабатываются (однако могут появиться в ответе, поскольку в тексте есть ссылки на них). Файлы других форматов игнорируются.

Создание интеграции

Чтобы загрузить тексты страниц с сайта:

1. Перейдите в раздел *Интеграции* и нажмите *Подключить → Веб-сайт*.

   Если при создании проекта вы выбрали *Подключить интеграцию*, то попадете на страницу новой интеграции автоматически.
2. Укажите настройки интеграции:



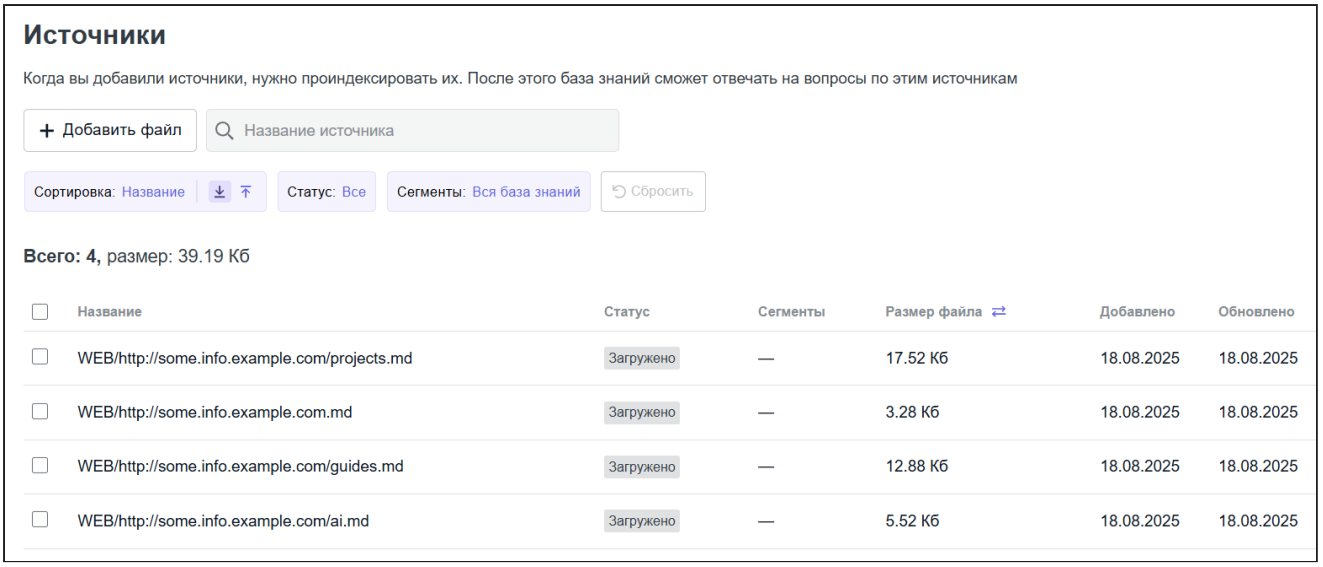
- *Название интеграции* — используется как префикс к названиям источников, загруженных из этой интеграции.
- *Сегменты* — к ним будут относиться источники, загруженные из этой интеграции. Сегменты позволяют сгруппировать источники с общей темой. В дальнейшем сегменты можно указывать в запросе к базе знаний для поиска информации только в этим сегментах. Подробнее читайте в разделе Сегменты.
- *Домен* — адрес сайта. Допустим поддомен любого уровня.
- *Выбрать ссылки для скачивания* — активируйте, чтобы вручную указать страницы для загрузки в базу знаний.
- *Автоматически синхронизировать данные* — включите для поддержания актуальности базы знаний, отключите для однократной загрузки.

> **⚠ ПРЕДУПРЕЖДЕНИЕ**
>
> Повторная обработка каждой страницы тарифицируется, даже если страница не менялась.

- *Автоматически перезапускать процесс индексации* — включите, чтобы переиндексировать базу знаний после очередной синхронизации.

> **ⓘ ПРИМЕЧАНИЕ**
>
> Все настройки впоследствии можно изменить.

Статус интеграции сразу после создания — «Идет обновление». Когда статус сменится на «Подключено», вы можете просмотреть загруженные страницы в разделе *Источники*.



В дополнение к данным из интеграций вы можете загрузить файлы вручную.

После загрузки данных нужно проиндексировать базу знаний.

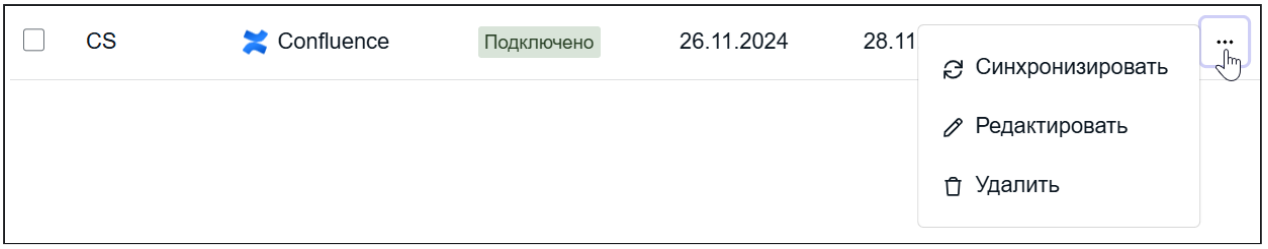
> **ⓘ К СВЕДЕНИЮ**
>
> О принудительной синхронизации и изменении настроек читайте в разделе Управление интеграциями.

Управление интеграциями

Принудительная синхронизация

Вручную запустить загрузку обновленных данных из интеграции вы можете в любой момент, независимо от настроек автоматической синхронизации.

1. Перейдите в раздел *Интеграции*.
2. Нажмите ⋯ напротив интеграции и выберите *Синхронизировать*.



3. Чтобы обновленные данные стали доступны пользователям, которые обращаются к базе знаний, запустите переиндексацию. Для этого нажмите *Индексировать* вверху страницы.

Изменение настроек

Чтобы изменить настройки интеграции, перейдите в раздел *Интеграции*, нажмите ⋯ напротив интеграции и выберите *Редактировать*.

Удаление интеграции

Чтобы удалить интеграцию и все ее данные из базы знаний:

1. Перейдите в раздел *Интеграции*, нажмите ⋯ напротив интеграции и выберите *Удалить*.
2. Перейдите в раздел *Источники* и удалите страницы, загруженные из этой интеграции:
   i. В поисковой строке введите название интеграции, которое служит префиксом к названиям страниц.
   ii. Выберите все найденные страницы и нажмите кнопку *Удалить* над списком источников.
3. Чтобы запустить переиндексацию, нажмите *Индексировать* вверху страницы.

> **ПРИМЕЧАНИЕ**

> Удаление отдельных источников из действующей интеграции нецелесообразно: они будут загружены обратно с очередным обновлением.

Индексация базы знаний

Перед использованием базы знаний нужно ее проиндексировать. Индексация базы знаний в Jay Knowledge Hub — это процесс, который включает в себя следующие шаги:

1. Подготовка загруженных данных.
2. **Чанкинг:** разделение данных на небольшие фрагменты (чанки).
3. **Векторизация:** преобразование текстовых фрагментов в векторы.

   Далее при подготовке ответа на запрос система будет сравнивать вектор запроса с векторами фрагментов базы знаний.

> ⓘ **К СВЕДЕНИЮ**
>
> Если вы хотите узнать больше о работе в Jay Knowledge Hub, переходите в справочный раздел документации.
>
> Параметры индексации можно изменить в разделе *Настройки проекта → Индексация*. Подробнее о настройках читайте в документации.

Как проиндексировать базу знаний

Нажмите *Индексировать* вверху любой страницы.

После этого начнется процесс индексации базы знаний, который может занять некоторое время в зависимости от объема данных. В это время вы можете покинуть страницу — это не повлияет на процесс индексации.

> ⚠ **ПРЕДУПРЕЖДЕНИЕ**
>
> Индексация произойдет на всех файлах и веб-страницах, которые были добавлены. Выбрать конкретные источники для индексации нельзя.
>
> Если вы хотите индексировать часть данных, удалите ненужные источники перед индексацией.

Когда индексация завершится, перейдите в раздел *Тестирование*, чтобы проверить работу вашей базы знаний.

Как переиндексировать базу знаний

Если вы загрузили новые данные или изменили настройки индексации, переиндексируйте базу знаний. Сам процесс переиндексации аналогичен индексации базы знаний.

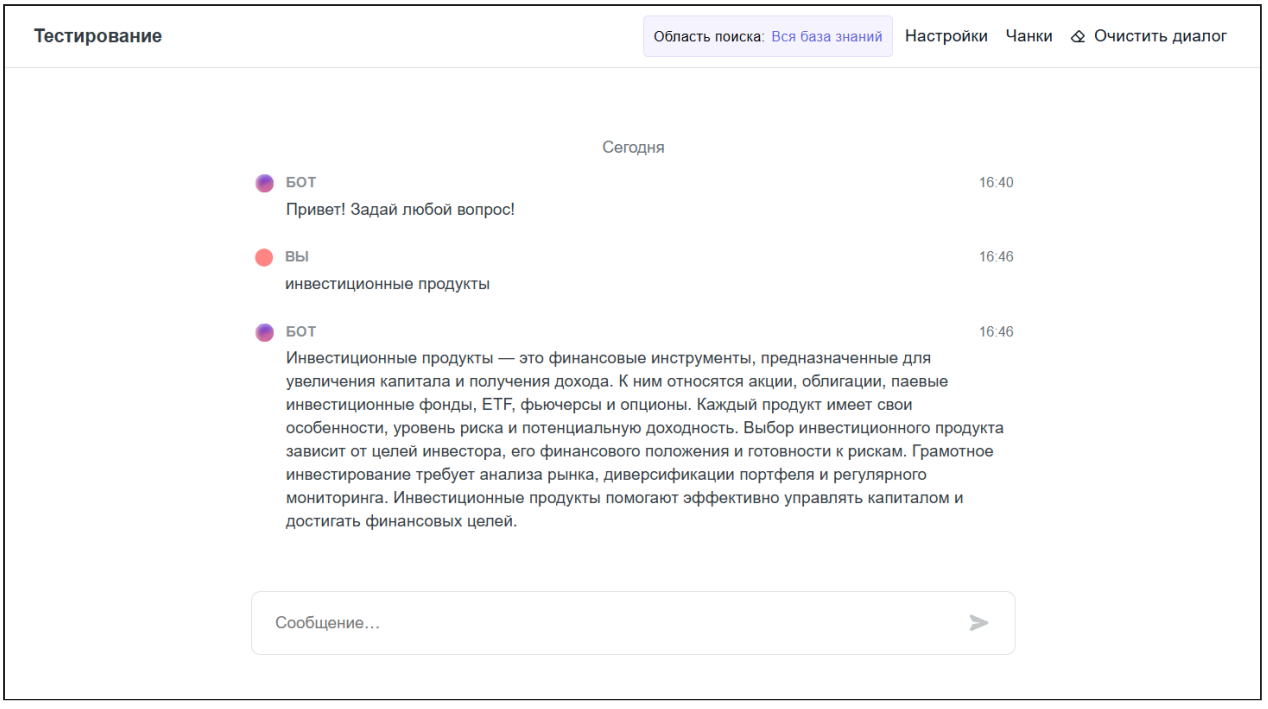
Учитывайте, что в случае изменения настроек предыдущий индекс будет удален. В частности, перестанут работать ссылки на изображения в старых ответах. Если канал, например мессенджер, не сохраняет изображения у себя, они исчезнут из диалога.

Тестирование базы знаний в интерфейсе

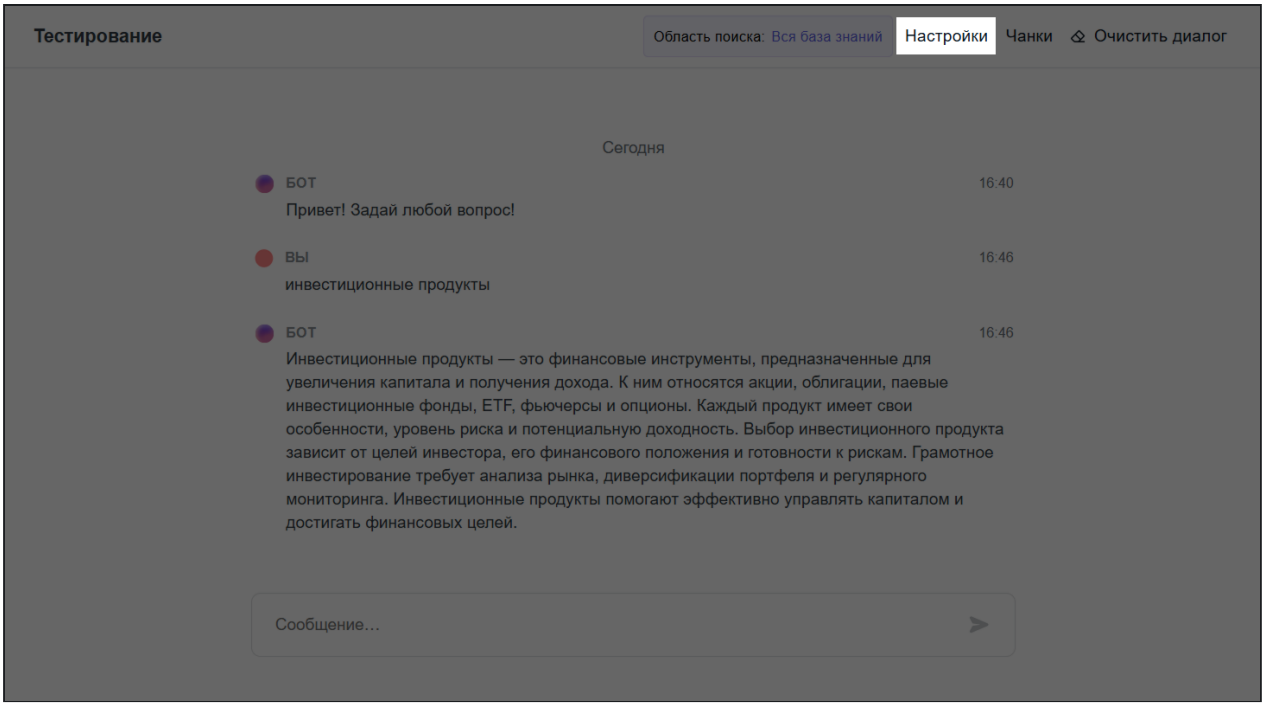
После того как база знаний проиндексирована, вы можете протестировать ее, чтобы убедиться, что она правильно работает и корректно отвечает на возможные вопросы пользователей.

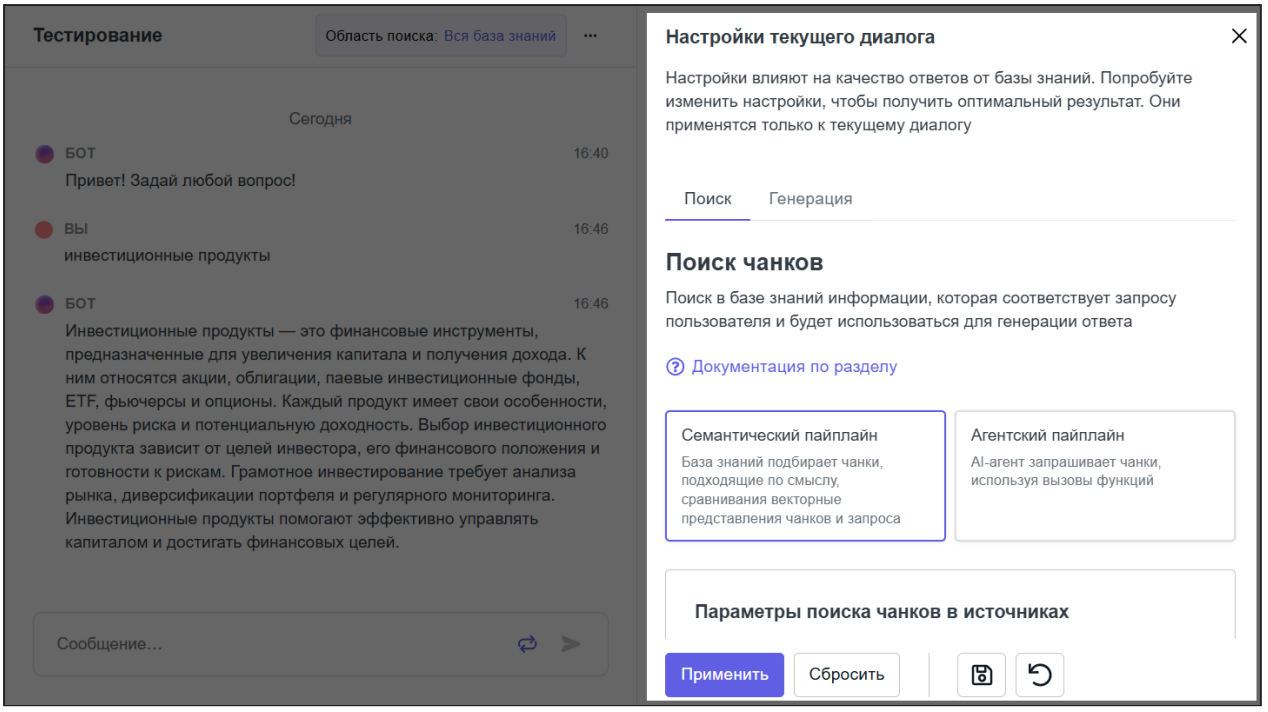
Как протестировать базу знаний в интерфейсе

1. Перейдите в раздел *Тестирование*.
2. Задавайте свои вопросы и смотрите, как база знаний отвечает на них. Попробуйте разные запросы, чтобы убедиться, что база знаний работает корректно:



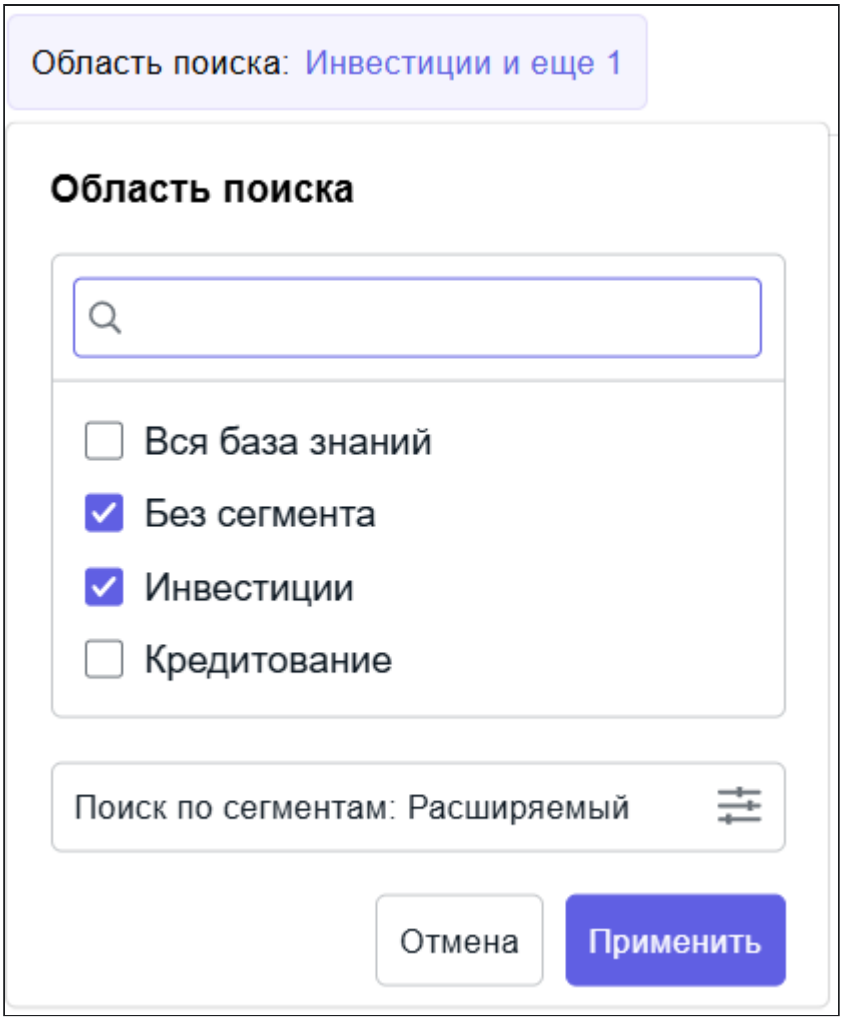
3. Перейдите в панель *Настройки*, чтобы попробовать другие параметры поиска и подготовки ответов:



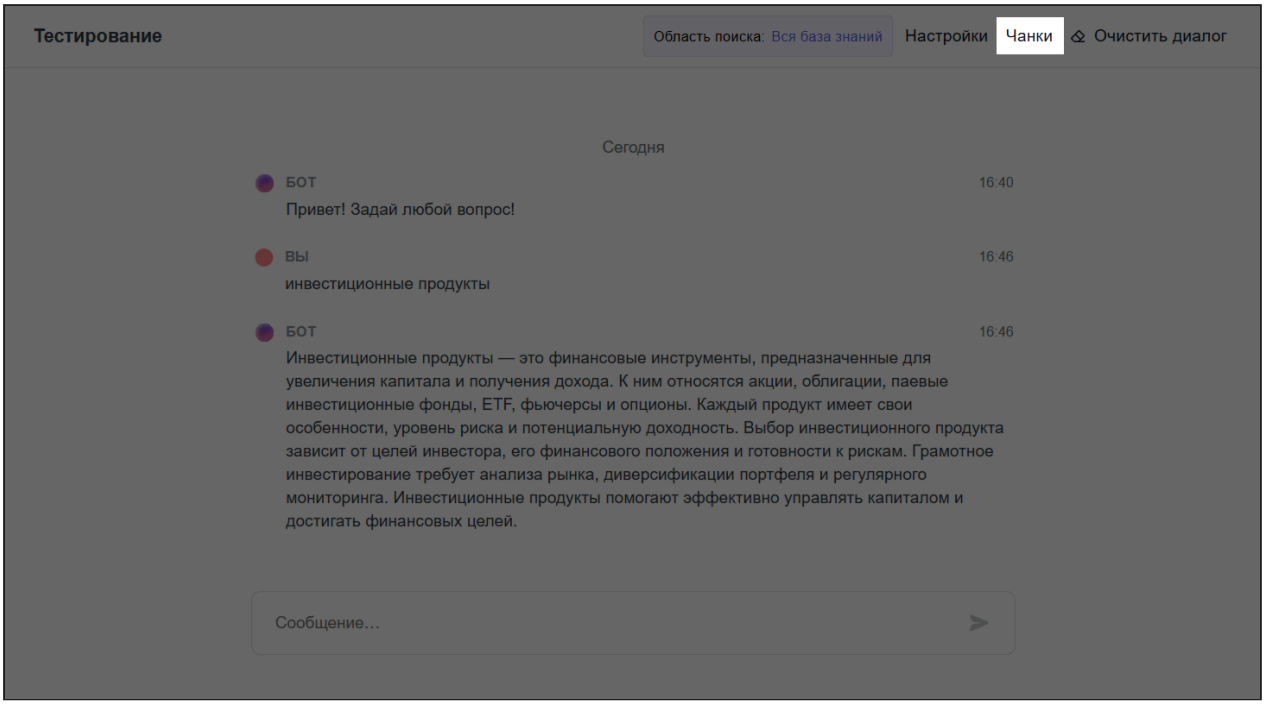


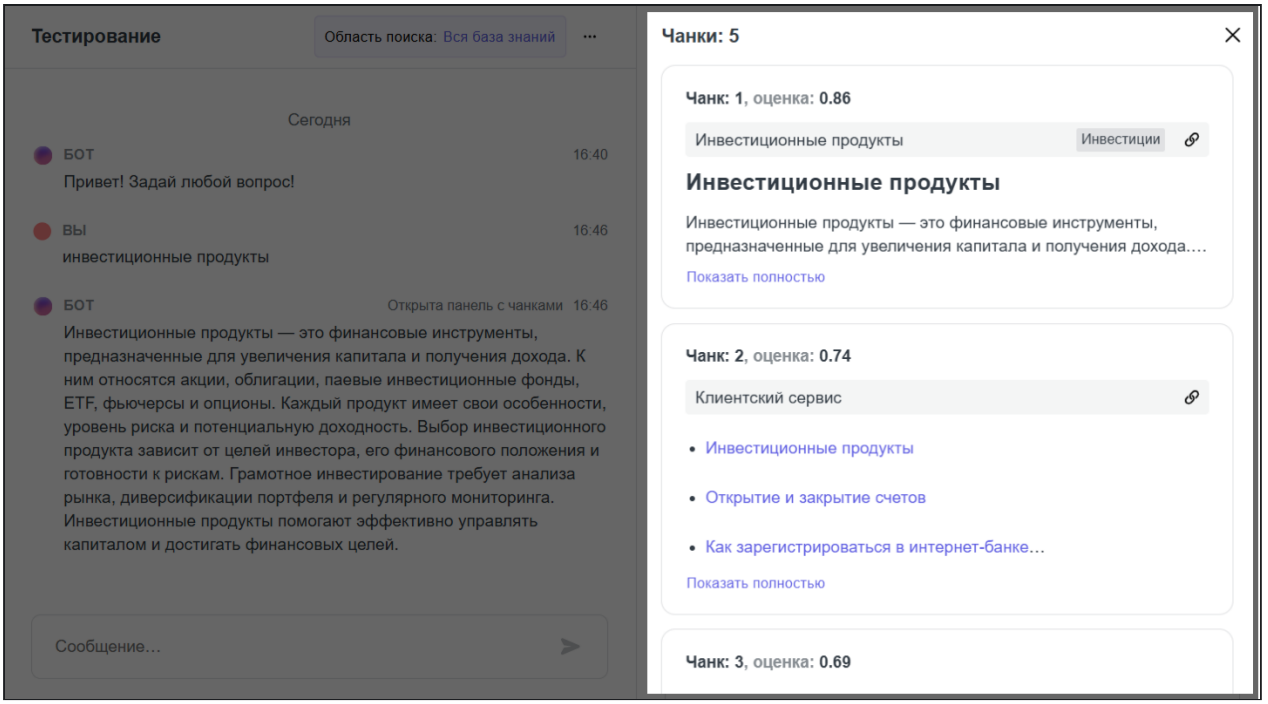
> **К СВЕДЕНИЮ**
>
> Эти настройки влияют только на текущий диалог. Чтобы применить их к базе знаний в целом, нажмите 💾. Подробнее о настройках читайте в документации.

4. Нажмите *Область поиска*, чтобы уточнить, в какой части базы знаний нужно искать информацию для ответа. Отметьте интересующие вас сегменты. Опция «Без сегмента» позволяет искать по источникам, которые не входят ни в один сегмент.



5. Перейдите в панель *Чанки*, чтобы посмотреть, какие фрагменты текста были использованы Jay Knowledge Hub для ответа на запрос:





ПОДСКАЗКА

Когда открыта одна из панелей *Настройки* или *Чанки*, используйте кнопку •••, чтобы открыть другую панель или очистить диалог.

Подключение каналов

Каналы — это интерфейсы, в которых ваши клиенты и сотрудники могут задавать вопросы к базе знаний: чат-виджеты на сайте, мессенджеры, соцсети и др.

Подключение каналов обеспечивает диалоговая платформа JAICP. Чтобы добавить каналы, необходимо предварительно создать проект JAICP, связанный с вашей базой знаний.

Подробнее о JAICP можно узнать на нашем сайте или в документации.

Создание связанного проекта JAICP

1. В базе знаний перейдите в раздел *Каналы*.
2. Нажмите кнопку *Создать проект JAICP*.

На платформе JAICP будет создан проект, связанный с вашей базой знаний и с таким же названием.

> **ПРИМЕЧАНИЕ**
>
> Чтобы создать проект JAICP и подключить каналы, необходима роль `SYSTEM_ADMIN` в Conversational Cloud. Если у вас нет этой роли, при попытке создать проект вы получите ошибку связи с JAICP. Чтобы получить роль, обратитесь к владельцу аккаунта.

Использование API-ключа

Для доступа к базе знаний проекту JAICP необходим API-ключ. Если в базе знаний еще нет API-ключей, система автоматически создает API-ключ с неограниченным сроком действия. Если есть — выбирает один из существующих API-ключей, с максимальным сроком действия. API-ключ для доступа к базе знаний записывается в проект JAICP в токен `khub_token`. Подробнее о токенах читайте в разделе Токены и переменные документации JAICP.

Если вы удалите API-ключ, используемый проектом JAICP, или истечет срок действия ключа, каналы потеряют доступ к базе знаний. В этом случае необходимо создать новый API-ключ, а затем в связанном проекте JAICP в разделе *Токены и переменные* заменить значение `khub_token`.

Подключение и настройка канала

К базе знаний можно подключить любой входящий канал, доступный в JAICP.

1. В базе знаний в разделе *Каналы* нажмите кнопку *Подключить канал в JAICP*.

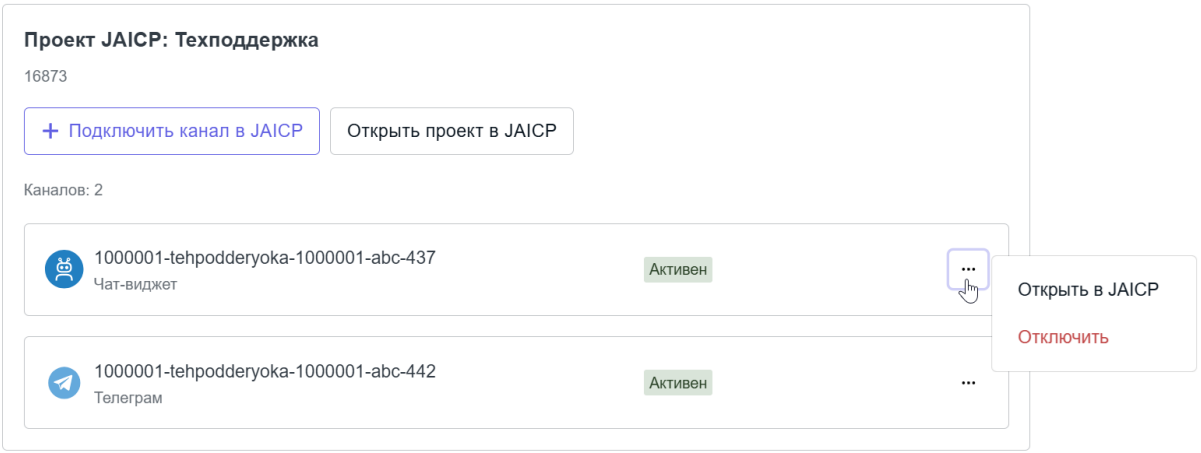
   В новой вкладке откроется связанный проект JAICP и раздел *Каналы*.
2. В блоке *Входящие* нажмите *Подключить канал*.
3. Выберите тип канала и укажите настройки.

> **ПОДСКАЗКА**
>
> Подробнее о каналах связи и их настройке читайте в разделе Каналы общения и интеграции документации JAICP.

Отключение канала

В базе знаний в разделе *Каналы* нажмите кнопку ••• для нужного канала и выберите *Отключить*.

В результате отключения запросы в этом канале обрабатываться не будут и останутся без ответа.



Подключение базы знаний в Jay Copilot

Jay Copilot от Just AI — это ассистент для применения генеративного искусственного интеллекта в работе и повседневных задачах. Jay Copilot предоставляет набор готовых приложений для маркетинга, продаж, HR, программирования, поддержки клиентов и других сфер.

Подробнее о Jay Copilot можно узнать на нашем сайте и в документации.

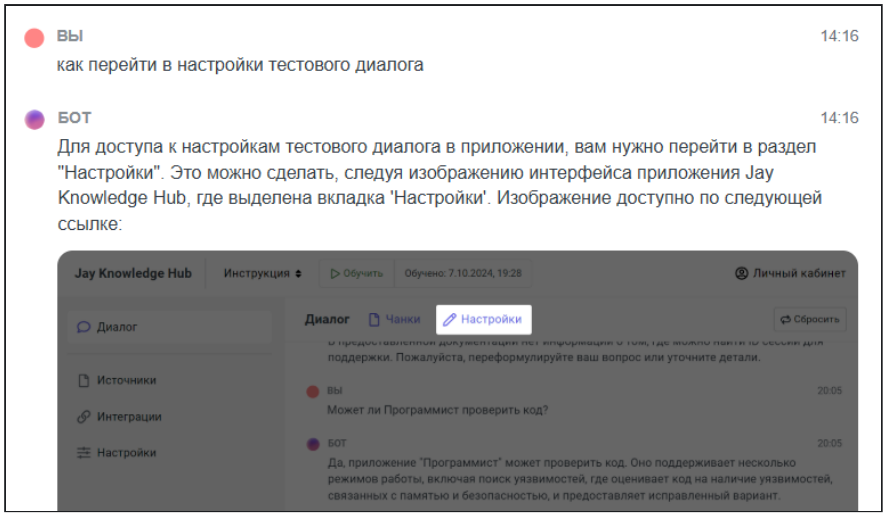
В Jay Copilot вы можете подключать свои базы знаний из Jay Knowledge Hub в качестве приложений. Обращаться к таким приложениям смогут все пользователи аккаунта.

Инструкцию по созданию приложения смотрите в разделе База знаний Jay Knowledge Hub документации Jay Copilot.

Изображения в ответе

Изображения из документов-источников могут появляться в ответах на вопросы пользователей.

По умолчанию поиск по базе знаний не учитывает содержание изображений — только текст вокруг них. Для поиска по изображениям включите в настройках проекта генерацию описаний изображений и переиндексируйте базу знаний. Эти описания будут использоваться для поиска и подготовки ответа наряду с текстом документа.



⚠ ПРЕДУПРЕЖДЕНИЕ

В каналах и в API база знаний возвращает временные ссылки на изображения, срок действия таких ссылок ограничен. Если канал, например мессенджер, не сохраняет изображения у себя, они исчезнут из диалога по истечении этого срока или после переиндексации базы знаний.

Оценка качества ответов базы знаний

Измерение качества ответов базы знаний позволяет решить несколько ключевых задач:

- Оценить, насколько эффективно база знаний удовлетворяет типичные запросы пользователей.
- Добиться максимально возможного качества ответа путем корректировки настроек и анализа результатов.
- Отслеживать изменения в качестве ответов с течением времени, по мере обновления источников данных.

Как проводится измерение

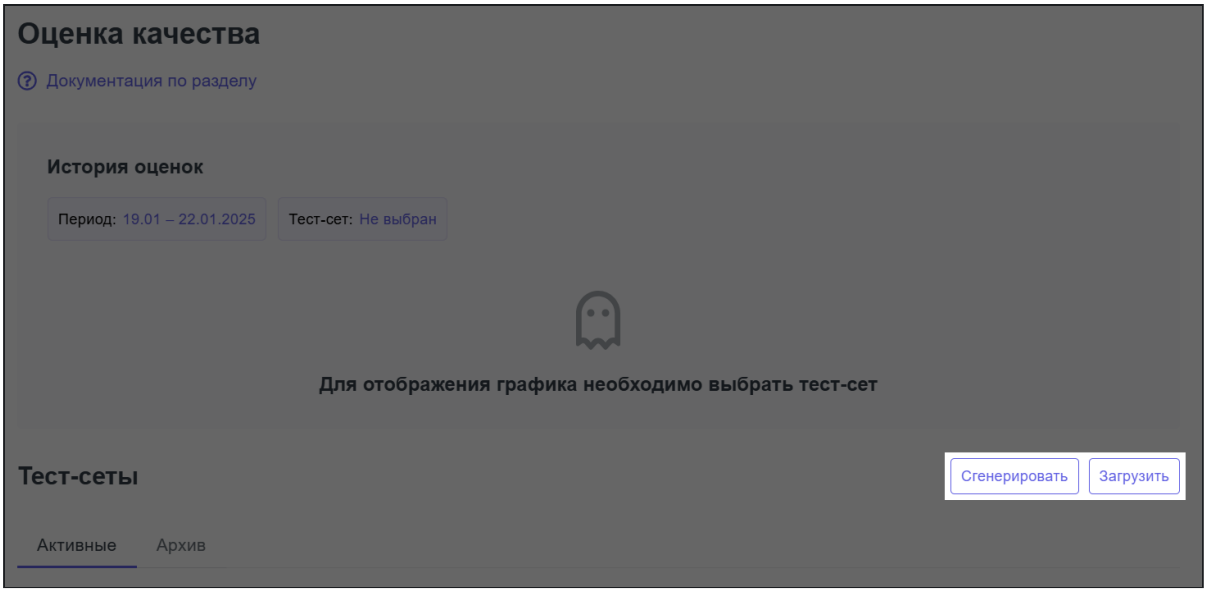
Для измерения используется тест-сет, который состоит из вопросов к базе знаний и ожидаемых ответов. Каждый вопрос отправляется в базу знаний, а полученный ответ вместе с вопросом и ожидаемым ответом передается в LLM. LLM оценивает качество фактического ответа по шкале от 1 до 10. Результат измерения — среднее значение по всем вопросам тест-сета.

Чем больше вопросов в тест-сете, тем достовернее результат измерения.

Качество ответа можно оценивать по нескольким тест-сетам. Измерения по разным тест-сетам проводятся независимо.

Подготовка тест-сета

Вы можете составить тест-сет самостоятельно или сгенерировать с помощью LLM.



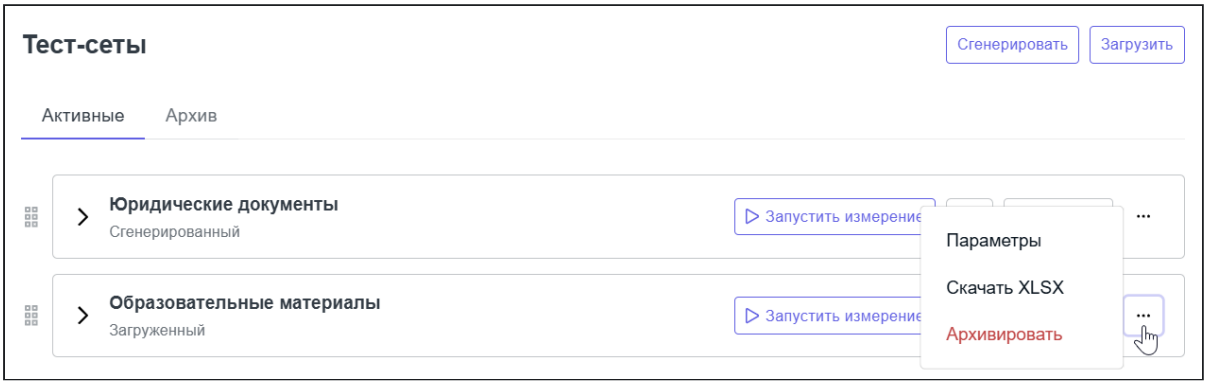
Самостоятельно **С помощью LLM**

1. Перейдите в раздел *Оценка качества*.
2. Скачайте шаблон тест-сета: в блоке *Тест-сеты* нажмите *Загрузить*, затем в окне загрузки нажмите *Скачать шаблон в XLSX*.
3. Внесите в файл вопросы к базе знаний и ожидаемые ответы.

   Если при загрузке источников или интеграции вы использовали сегменты, вы можете указать их для вопросов в столбце *Segments*. В этом случае для ответа на вопрос поиск будет идти только по источникам с соответствующим сегментом.

   Чтобы указать несколько сегментов, перечислите их через запятую. Для источников без сегментов укажите `include_without_segments`.
4. В окне загрузки укажите название тест-сета, которое будет отображаться в списке тест-сетов, и прикрепите заполненный файл.
5. Нажмите *Создать*.

В блоке *Тест-сеты* вы можете скачать загруженный тест-сет.



Запуск измерения

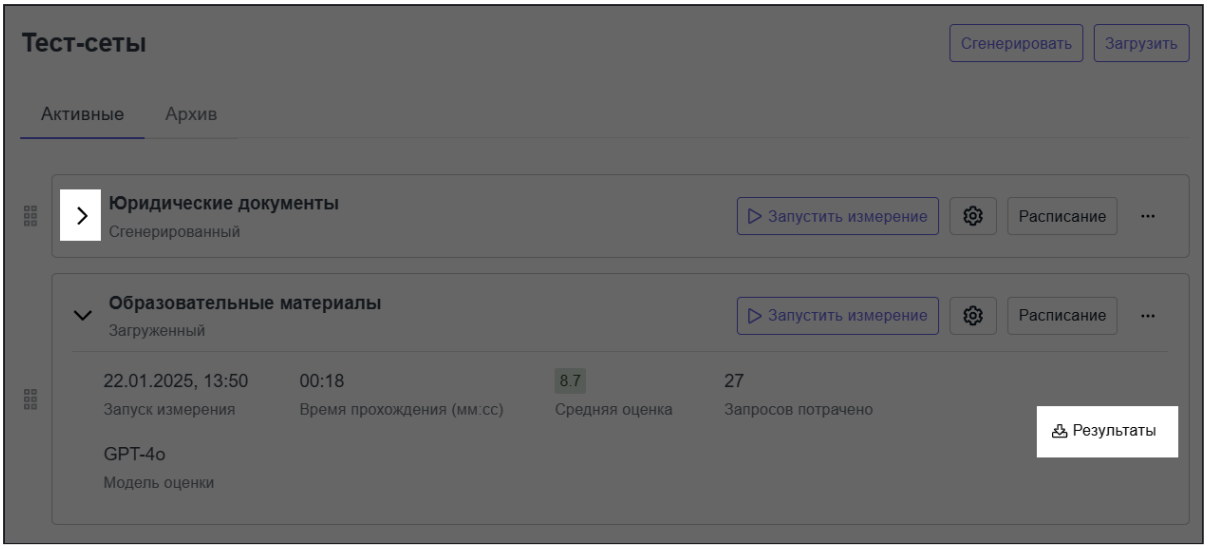
Чтобы указать модель для оценки качества ответов, в строке с тест-сетом нажмите ⚙. Выберите модель из списка и нажмите *Сохранить и запустить*.

Если модель уже была выбрана, в строке с тест-сетом нажмите *Запустить измерение*.

Измерение может занять длительное время, до нескольких часов, в зависимости от количества вопросов в тест-сете.

Просмотр результатов измерения

Чтобы скачать подробный отчет с оценками по каждому вопросу из тест-сета, нажмите › в строке с тест-сетом, а затем нажмите *Результаты* для интересующего вас измерения.



Настройка расписания

Чтобы запускать измерение с определенной периодичностью:

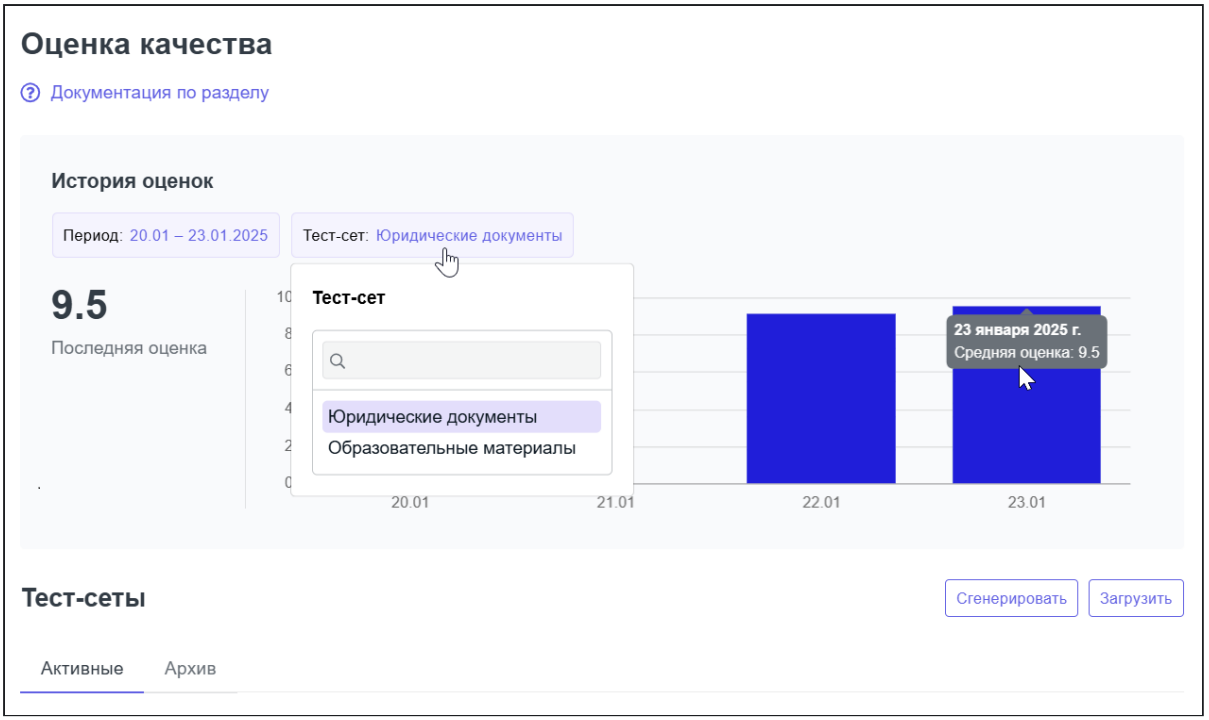
1. В строке с тест-сетом нажмите *Расписание*.
2. Укажите периодичность и время запуска.
3. Чтобы пропускать очередное измерение по расписанию, если в базе знаний не было обновлений с предыдущего измерения, используйте опции в блоке *Измерять только при обновлениях*:
   - *В источниках* — запускать измерение по расписанию, только если были обновлены данные в базе знаний.
   - *В настройках* — запускать измерение по расписанию, только если были обновлены настройки проекта.

- Включите обе опции, чтобы запускать измерение по расписанию при наличии любых обновлений.

Просмотр истории измерений

История оценок доступна по каждому тест-сету отдельно.

Чтобы посмотреть график, в блоке *История оценок* выберите нужный период и тест-сет.



Чтобы посмотреть список измерений, в строке с тест-сетом нажмите ›.

Тарификация

Для создания и использования базы знаний требуются **J-токены**. J-токен — это единица тарификации в Jay Knowledge Hub, общая для всех типов расходов.

При регистрации вы бесплатно получаете 1 000 J-токенов. Когда токены закончатся, приобретите пакет токенов, чтобы продолжать пользоваться базой знаний. Отслеживать и пополнять баланс токенов можно на странице *Личный кабинет*.

> ⚠ **ПРЕДУПРЕЖДЕНИЕ**
>
> Актуальная информация о ценах отражена только на этой странице. Базы знаний не могут дать точную информацию о том, сколько стоит их использование.

Пакеты J-токенов

Размер пакета	Стоимость
500 J-токенов	500 ₽
1 000 J-токенов	1 000 ₽
5 000 J-токенов	5 000 ₽

Для компаний предусмотрены корпоративные тарифы: чем больше пакет J-токенов, тем выгоднее. Если вам это интересно, свяжитесь с нами: contact@just-ai.com.

Расход J-токенов

Посмотреть, как расходуются J-токены, вы можете в проекте базы знаний в разделе *Статистика*.

Запросы к LLM

База знаний отправляет в LLM запросы нескольких типов:

- Пользовательские запросы — для перефразирования вопроса пользователя к базе знаний и для генерации ответа. Учитываются вопросы, заданные в тестовом диалоге, в каналах и через API.
- Агентские запросы — для поиска чанков, если в настройках выбран метод поиска чанков с помощью LLM.
- Оценка качества — для генерации тест-сетов и проведения измерений.
- Чанкинг:
  - для разделения текста на чанки, если выбран чанкинг с помощью LLM;
  - для обогащения чанков дополнительной информацией и получения описаний изображений, если эти опции включены в настройках.

Стоимость запроса в J-токенах зависит:

- от выбранной языковой модели;
- от длины запроса (вместе с контекстом) и ответа на него.

Модель	**Стоимость 1 000 токенов модели, в J-токенах**	
	В запросе ↑	**В ответе ↓**
DeepSeek-V3	0,27	1,1
GigaChat Lite	1	1
GigaChat Pro	7,5	7,5
Qwen2.5-32B-Instruct-fp8-dynamic	0,625	0,625
Qwen/Qwen2.5-72B-Instruct-AWQ	0,875	0,875
T-Lite	0,7	0,7
T-Pro	1,25	1,25
YandexGPT Lite	1	1
YandexGPT Pro	3	3

Хранение данных

Стоимость хранения документов-источников — 0,56 J-токенов в день за гигабайт (16,9 J-токенов в месяц).

Если баланс J-токенов исчерпан, файлы хранятся еще 30 дней, а затем удаляются.

Векторизация

Векторизация — преобразование текста в числовые векторы. Применяется как к вашим данным на этапе индексации, так и к запросам пользователей.

Стоимость зависит от выбранной модели векторизатора:

- `text-embedding-3-large` — 0,13 J-токенов за 1 000 токенов модели.
- `intfloat/multilingual-e5-large` — 0,1 J-токена за 1 000 токенов модели.

Парсинг сайта

Стоимость обработки одной страницы из интеграции с веб-сайтом — 11 J-токенов.

Приглашение пользователей

Jay Knowledge Hub поддерживает совместную работу в аккаунте.

Доступ для новых пользователей

Вы можете приглашать пользователей в те аккаунты, которыми управляете. Для таких аккаунтов отображается надпись «♔ Управляю».

Чтоб пригласить пользователей:

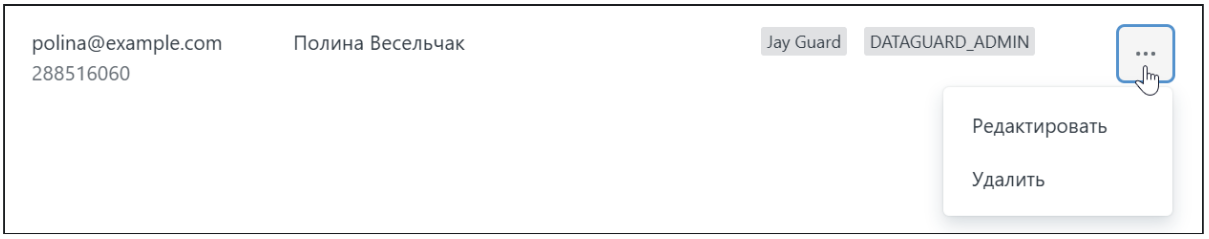
1. Перейдите в раздел Аккаунты в Conversational Cloud.
2. Выберите аккаунт, которым вы управляете.
3. В аккаунте нажмите *Пригласить пользователя*.
4. Выберите продукт Knowledge Hub.
5. Выберите группы и роли для пользователей:
   - В группе *Knowledge Hub* доступна единственная роль — KHUB_ADMIN. Она предоставляет полный доступ к Jay Knowledge Hub.
   - В группе *General* актуальная роль для Jay Knowledge Hub — USER_ADMIN. Она позволяет просматривать пользователей аккаунта и управлять ими, например приглашать новых пользователей и изменять роли существующих.
6. Укажите email пользователей.
7. Нажмите *Пригласить*.

Подробно о приглашении пользователей читайте в документации Conversational Cloud.

Доступ для ранее приглашенных пользователей

Пользователям вашего аккаунта, у которых еще нет доступа к Jay Knowledge Hub, вы можете предоставить его:

1. Перейдите в раздел Аккаунты в Conversational Cloud.
2. Выберите аккаунт, которым вы управляете.
3. В аккаунте нажмите ⋯ напротив пользователя и выберите *Редактировать*.

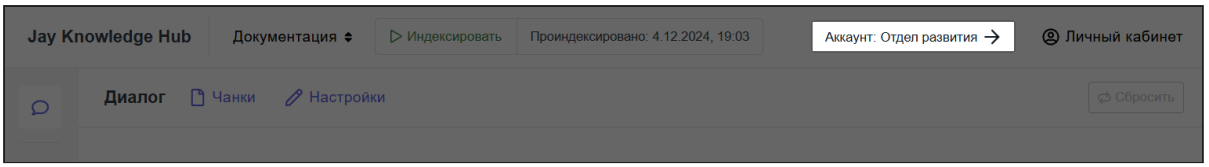


4. Выберите группу и роль KHUB_ADMIN.
5. Нажмите *Сохранить*.

Переключение между аккаунтами

Если вы владелец или участник нескольких аккаунтов, вы можете переключаться между ними:

1. Нажмите кнопку с названием аккаунта (вверху справа на любой странице).



2. В разделе *Ваши аккаунты* выберите нужный аккаунт.

> **ПРИМЕЧАНИЕ**
>
> Если кнопка переключения аккаунтов не отображается, значит, в других аккаунтах у вас нет доступа к Jay Knowledge Hub. Для получения доступа обратитесь к владельцу аккаунта.

API Jay Knowledge Hub

Jay Knowledge Hub предоставляет API для взаимодействия с базой знаний. API позволяет интегрировать базу знаний с сайтами и приложениями, чтобы ваши клиенты и сотрудники могли обращаться к ней с вопросами.

> **ПОДСКАЗКА**
>
> Методы, параметры запросов, форматы ответов, возвращаемые ошибки подробно описаны в спецификации API.

Доступ к API

В Jay Knowledge Hub предусмотрено два типа API-ключей:

- **Административный** — позволяет добавлять и удалять проекты, создавать и отзывать и API-ключи проектов.
- **Проектный** — предоставляет доступ к базе знаний проекта: позволяет выполнять запросы к ней, управлять источниками и интеграциями.

Чтобы создать API-ключ:

1. Перейдите в *Личный кабинет → API-ключи*.
2. Нажмите *Создать API-ключ*.
3. Выберите тип ключа: проектный или административный.
4. Для проектного ключа выберите проект.
5. Введите название ключа.
6. Если требуется, укажите срок действия.

API-ключ нужно указывать в каждом запросе к API в HTTP-заголовке:

```
Authorization: Bearer <your_API_key>
```

Базовый URL

```
```

```
https://khub.just-ai.com/api/knowledge-hub
```

Список методов

API включает набор методов для взаимодействия с вашими знаниями:

- Создание, изменение, удаление проектов.
- Создание и деактивация проектных API-ключей.
- Получение информации о проекте базы знаний, включая список доступных LLM.
- Добавление, обновление и удаление источников.
- Добавление и удаление интеграций для загрузки данных из внешних систем.
- Запуск и остановка индексации базы знаний.
- Поиск релевантных чанков для запроса пользователя.
- Генерация ответов на запросы пользователей.
- Создание и поддержание чатов с пользователем.

Проекты

Метод	Назначение	Ключ*	Документация
`POST /projects`	Создание проекта.	А	Подробнее
`GET /projects`	Получение списка проектов.	А	Подробнее
`PUT /projects/{projectId}`	Обновление проекта.	А	Подробнее
`DELETE /projects/{projectId}`	Удаление проекта.	А	Подробнее
`GET /info`	Получение информации о проекте.	П	Подробнее

* А — для вызова метода необходим административный API-ключ, П — проектный.

API-ключи

Для вызова методов необходим административный API-ключ.

Метод	Назначение	Документация
`POST /projects/{projectId}/api-keys:`	Создание проектного API-ключа.	Подробнее
`GET /projects/{projectId}/api-keys`	Получение списка API-ключей проекта.	Подробнее
`POST /projects/{projectId}/api-keys/{apiKeyId}/revoke`	Деактивация проектного API-ключа.	Подробнее

Источники

Для вызова методов необходим проектный API-ключ.

Метод	Назначение	Документация
`GET /sources`	Получение списка источников.	Подробнее
`POST /sources/links`	Добавление источника по ссылке.	Подробнее
`PUT /sources/links`	Обновление источника по ссылке.	Подробнее
`POST /sources/texts`	Добавление текста в качестве источника.	Подробнее
`PUT /sources/texts`	Обновление источника из текста.	Подробнее
`POST /sources/files`	Добавление файла в качестве источника.	Подробнее
`PUT /sources/files`	Обновление источника из файла.	Подробнее
`GET /sources/{sourceId}`	Получение информации об источнике.	Подробнее

Метод	Назначение	Документация
`GET /sources/{sourceId}/download`	Выгрузка источника.	Подробнее
`DELETE /sources/{sourceId}`	Удаление источника.	Подробнее

Интеграции

Для вызова методов необходим проектный API-ключ.

Метод	Назначение	Документация
`GET /integrations`	Получение списка интеграций.	Подробнее
`POST /integrations`	Добавление интеграции.	Подробнее
`GET /integrations/{integrationId}`	Получение информации об интеграции.	Подробнее
`DELETE /integrations/{integrationId}`	Удаление интеграции.	Подробнее

Сегменты

Метод	Назначение	Документация
`GET /segments`	Получение списка сегментов.	Подробнее
`POST /segments/rename`	Переименование сегмента.	Подробнее
`POST /sources/segments`	Привязка сегментов к источникам.	Подробнее
`DELETE /sources/segments`	Отвязка сегментов от источников.	Подробнее

Индексация

Для вызова методов необходим проектный API-ключ.

Метод	Назначение	Документация
`GET /indexing/start`	Запуск индексации базы знаний.	Подробнее
`GET /indexing/stop`	Остановка индексации базы знаний.	Подробнее

Одиночные запросы

Обработка запросов на поиск чанков и генерацию ответов. Дополнительно можно передать в запросе историю сообщений.

Для вызова методов необходим проектный API-ключ.

Метод	Назначение	Документация
`POST /retrieve`	Поиск чанков (retrieving).	Подробнее
`POST /query`	Генерация ответа (синхронный вызов).	Подробнее
`POST /async/query`	Генерация ответа (асинхронный вызов).	Подробнее
`GET /query/{queryId}`	Получение статуса генерации ответа.	Подробнее
`POST /query/{queryId}/cancel`	Отмена генерации ответа.	Подробнее

Чаты и запросы в рамках чата

Создание чата, обработка запросов на поиск чанков и генерацию ответов в рамках чата. При этом учитывается история сообщений в чате.

Для вызова методов необходим проектный API-ключ.

Метод	Назначение	Документация
`POST /chat`	Создание чата.	Подробнее

Метод	Назначение	Документация
`GET /chat/{chatId}`	Получение информации о чате.	Подробнее
`POST /chat/{chatId}/retrieve`	Поиск чанков (retrieving).	Подробнее
`POST /chat/{chatId}/query`	Генерация ответа (синхронный вызов).	Подробнее
`POST /async/chat/{chatId}/query`	Генерация ответа (асинхронный вызов).	Подробнее
`GET /chat/{chatId}/query/{queryId}`	Получение статуса генерации ответа.	Подробнее
`POST /chat/{chatId}/query/{queryId}/cancel`	Отмена генерации ответа.	Подробнее

Варианты использования

Одиночный запрос

В запросе можно указать настройки обработки запроса. По умолчанию используются настройки проекта. Дополнительно можно передать в запросе историю сообщений.

Пример запроса:

```
curl \
--header 'Authorization: Bearer <your_API_key>' \
--header 'Content-Type: application/json' \
--data '{
    "query": "Как создать собственную базу знаний"
}' \
https://khub.just-ai.com/api/knowledge-hub/query
```

▼ Пример ответа

```
{
    "id": 20199,
    "request": "Как создать собственную базу знаний",
    "status": "FINISHED",
    "createdAt": "2024-11-18T12:01:56.166692Z",
    "response": "Для создания собственной базы знаний в Jay Knowledge Hub выполните следующие шаги:\n\n1. **Регистрация**: Зарегистрируйтесь в Jay Knowledge Hub. Если у вас уже есть учетная запись от других продуктов Just AI, вы можете использовать её для входа на платформу.\n\n2. **Создание проекта и загрузка данных**: Создайте ваш первый проект и загрузите данные в Jay Knowledge Hub. Поддерживаются файлы форматов PDF, DOCX, TXT, XML, JSON, YAML, MD, MDX, HTML. Если вы используете Confluence для хранения знаний, настройте интеграцию с Jay Knowledge Hub для автоматической загрузки данных.\n\n3. **Индексация базы знаний**: Запустите процесс индексации загруженных данных. Индексация может занять некоторое время, следите за прогрессом в интерфейсе продукта.\n\nПосле выполнения этих шагов ваша база знаний будет готова к использованию.",
    "updatedAt": "2024-11-18T12:02:04.481433Z",
    "comment": null
}
```


Чат

1. Создайте чат.

   В запросе можно указать название чата и настройки обработки запросов в чате.

   По умолчанию используются настройки проекта.

   Пример запроса:

```
curl \
--header 'Authorization: Bearer <your_API_key>' \
--header 'Content-Type: application/json' \
--data '{}' \
https://khub.just-ai.com/api/knowledge-hub/chat
```

   ▼ Пример ответа


```
{
    "id": 15045,
    "settings": {
        "pipeline": "semantic",
        "search": {
            "similarityTopK": 5,
            "candidateRadius": 0,
            "reranker": null,
```

```
            "fullTextSearch": null,
            "rephraseUserQuery": null,
            "segment": null
        },
        "llm": {
            "model": "yandexgpt",
            "contextWindow": 16000,
            "maxTokens": 1000,
            "temperature": 0.0,
            "topP": 1.0,
            "frequencyPenalty": 0.0,
            "presencePenalty": 0.0
        },
        "responseGeneration": {
            "prompt": "You are a Russian-speaking question answering system.
You help users look for information in the Documentation.\nTry to answer their
Question or find documents related to the Question according to the
Documentation provided below.\nEach document starts from New document.\nFollow
the rules:\n- (1) Firstly, find relevant documents related to the Question. It
is possible that there are no related documents in the Documentation.\n- (2)
Answer strictly according to the documents. Don't make things up.\n- (3)
Question can be formulated as either an ordinary question with a question mark
or words (когда работает банк?) or a search query (режим работы банка). You
must deal with all types.\n- (4) If there is more than one answer to the
Question, list all possible solutions separately.\n- (5) If there is not a
direct Answer, but just information on the related topic, answer with it.\n-
(6) If you can't find relevant documents, ask a user to reformulate the
Question in your answer.\n- (7) Answer the Question in Russian.\n- (8) If you
see a URL to an image in Documentation, use it in your response.\n\nYour
task:\n_____\n\nDOCUMENTATION: <{context_str}>;\nQUESTION:
{query_str};\nANSWER:"
        }
    },
    "name": "2024-11-18-13-47-06"
}
```


2. Отправьте запрос к базе знаний в асинхронном режиме.

   Укажите идентификатор чата, созданного на предыдущем шаге, и пользовательский запрос. Дополнительно можно указать настройки обработки запроса. По умолчанию используются настройки обработки запросов в данном чате.

   Пример запроса:

```
curl \
--header 'Authorization: Bearer <your_API_key>' \
--header 'Content-Type: application/json' \
--data '{
    "query": "Как создать собственную базу знаний"
```

```
}' \
https://khub.just-ai.com/api/knowledge-hub/async/chat/15045/query
```

Пример ответа


```
{
    "id": 17748,
    "chatId": 15045,
    "request": "Как создать собственную базу знаний",
    "status": "READY_TO_PROCESS",
    "createdAt": "2024-11-18T13:59:17.642584567Z",
    "response": null,
    "updatedAt": "2024-11-18T13:59:17.642585762Z",
    "comment": null
}
```


3. Отслеживайте готовность ответа.

   В запросе укажите идентификатор чата и идентификатор запроса, полученный в результате асинхронного вызова. Дополнительно можно указать параметр `waitTimeSeconds`, чтобы использовать long polling. Когда генерация завершена, возвращается статус `FINISHED`.

   Пример запроса:

```
curl \
--header 'Authorization: Bearer <your_API_key>' \
https://khub.just-ai.com/api/knowledge-hub/chat/15045/query/17748
```

Пример ответа


```
{
    "id": 17748,
    "chatId": 15045,
    "request": "Как создать собственную базу знаний",
    "status": "FINISHED",
    "createdAt": "2024-11-18T13:59:17.642585Z",
    "response": "Для создания собственной базы знаний в Jay Knowledge Hub
следуйте следующим шагам:\n\n1. **Регистрация**: Зарегистрируйтесь в Jay
Knowledge Hub. Если у вас уже есть учетная запись от других продуктов Just AI,
вы можете использовать её для входа на платформу.\n\n2. **Создание проекта и
загрузка данных**: Создайте ваш первый проект и загрузите данные в Jay
Knowledge Hub. Поддерживаются файлы форматов PDF, DOCX, TXT, XML, JSON, YAML,
MD, MDX, HTML, CSV. Если вы используете Confluence для хранения знаний,
```

```
настройте интеграцию с Jay Knowledge Hub для автоматической загрузки данных.\n\n3. **Индексация базы знаний**: Запустите процесс индексации базы знаний на загруженных данных. Индексация может занять некоторое время, следите за прогрессом в интерфейсе продукта.\n\n4. **Тестирование и использование**: После индексации протестируйте базу знаний, чтобы убедиться в её корректной работе. Перейдите в раздел \"Тестирование\" и задавайте вопросы, чтобы проверить, как база знаний отвечает на них.\n\nЭти шаги помогут вам создать и настроить собственную базу знаний в Jay Knowledge Hub.",
    "updatedAt": "2024-11-18T13:59:26.181718Z",
    "comment": null
}
```

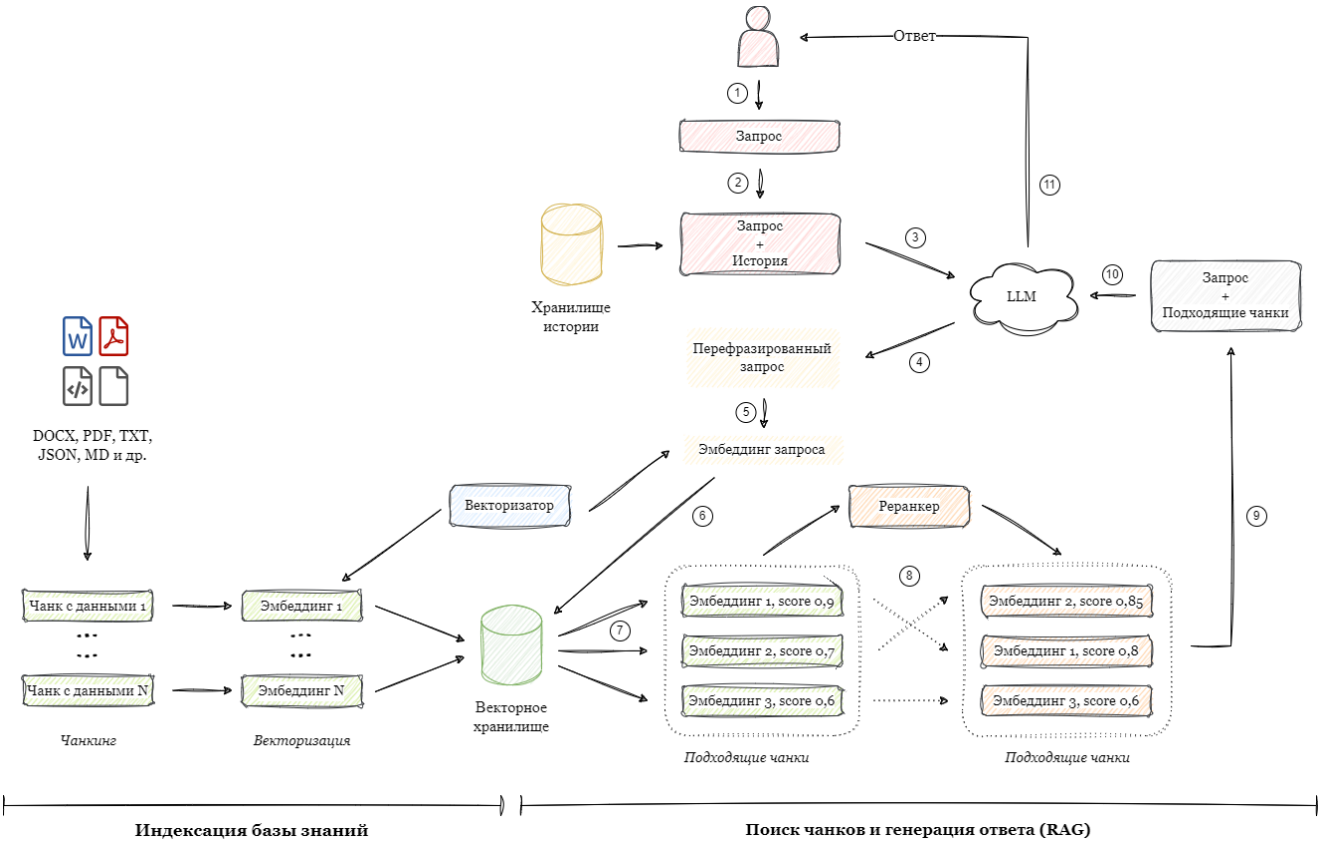
Основные этапы работы Jay Knowledge Hub

В основе работы Jay Knowledge Hub лежит механизм RAG (Retrieval-Augmented Generation), который сочетает в себе поиск и генерацию ответов на запросы пользователей. RAG объединяет внутренние знания LLM с обширными и постоянно обновляющимися внешними источниками данных, такими как базы данных, веб-сайты и другие информационные ресурсы.

С помощью RAG можно повысить точность и достоверность генерации ответов, особенно для задач, требующих глубоких знаний. Механизм также позволяет постоянно обновлять знания и интегрировать информацию, специфичную для конкретной области.

В зависимости от настроек проекта выполняется один из следующих пайплайнов.

Семантический пайплайн | **Агентский пайплайн**



Индексация базы знаний

Для индексации базы знаний используются различные источники данных: ваши файлы в форматах DOCX, PDF, TXT и др., а также внешние сервисы хранения знаний, такие как Confluence. Данные из источников проходят предварительную обработку, в результате которой получаются MD-файлы с текстом, которые затем передаются на чанкинг и векторизацию.

Чанкинг

Чанкинг — это разделение текста на небольшие фрагменты (чанки), которые затем используются для поиска и подготовки ответов на запросы пользователей.

Кроме того, чанкинг может включать подготовку данных с помощью LLM, если это указано в настройках проекта:

- Обогащение чанков дополнительной информацией для улучшения качества поиска.

  Дополнительная информация — это заголовок, краткое содержание, ключевые слова и вопросы, на которые отвечает чанк. Часть этой информации добавляется непосредственно в контент чанка, а часть в виде отдельных полей — метаданных чанка.
- Генерация описаний изображений (исключая иконки).

Векторизация

Векторизация — это преобразование текста в векторное представление. Каждый чанк преобразуется в вектор, который затем используется для поиска наиболее подходящего ответа на запрос пользователя. Векторизация позволяет использовать геометрические и алгебраические операции для сравнения и анализа данных.

Векторизованные чанки (эмбеддинги) сохраняются в векторном хранилище, оптимизированном для быстрого поиска и доступа к данным.

Retrieval-Augmented Generation (RAG)

Семантический пайплайн | **Агентский пайплайн**

Retrieving

Запрос пользователя перефразируется с учетом истории диалога. Перефразированный запрос так же, как и чанки с данными, преобразуется в векторное представление. Поиск информации выполняется путем сравнения векторных представлений запроса пользователя и данных.

Векторное хранилище автоматически подбирает наиболее подходящие данные, основываясь на сходстве векторов.

Повторное ранжирование (реранкинг)

После того как векторное хранилище нашло наиболее подходящие чанки, они проходят повторное ранжирование с помощью специальной модели — реранкера. При этом оценка релевантности каждого чанка может поменяться по сравнению с оценкой, полученной на этапе подбора векторным хранилищем.

Генерация ответа

На последнем этапе система генерирует ответ на запрос пользователя. Система отправляет в LLM запрос и чанки, отобранные на предыдущих этапах. Затем LLM генерирует ответ, используя предоставленную информацию, а также свои внутренние знания.

Настройки базы знаний

Чтобы влиять на поведение базы знаний, используйте настройки. Настройки позволяют вам сконфигурировать процесс индексации данных, а также поиск чанков и генерацию ответов.

Чтобы перейти в раздел настроек, используйте панель навигации слева.

Читать далее

- Настройки индексации
- Настройки поиска чанков
- Настройки генерации ответов

Настройки индексации

Индексация данных, загруженных в базу знаний, состоит из нескольких этапов:

1. Обработка данных — преобразование текста в формат MD.
2. Чанкинг — разбиение текста на фрагменты (чанки).
3. Векторизация — преобразование полученных чанков в векторные представления (эмбеддинги).

В разделе *Настройки проекта → Индексация* вы можете менять параметры для чанкинга и векторизации.

> ⚠ **ПРЕДУПРЕЖДЕНИЕ**
>
> После изменения настроек индексации нужно будет заново проиндексировать базу знаний.

Векторизация

Параметр *Модель векторизатора* определяет языковую модель для векторизации текста. Эта модель будет применяться для векторизации как ваших данных, так и запросов от пользователей:

- *text-embedding-3-large* — модель от компании OpenAI. При ее использовании ваши данные будут отправляться на зарубежный сервер.
- *intfloat/multilingual-e5-large* — модель, размещенная на серверах компании Just AI, которые находятся в России.

Чанкинг

Способ разделения текста

Параметр *Способ разделения текста* определяет, каким образом текст будет разбит на чанки:

- *По длине* — текст будет разбит на части по длине, с учетом границ слов.
- *С помощью LLM* — текст будет разбит на части с помощью языковой модели. В этом случае чанки будут учитывать иерархическую структуру текста: заголовки, абзацы, названия разделов и документа.

Состав настроек зависит от выбранного способа разделения.

По длине **С помощью LLM**

- *Максимальный размер чанка, в символах*.

  ▼ Как будет разбит текст

  Допустим, в настройке указано значение 70. У вас есть текст из 2 предложений, каждое из которых по 100 символов.

  При чанкинге текст будет разбит на 3 чанка:

  i. 70 символов первого предложения.
  ii. Оставшиеся 30 символов первого предложения и 40 второго.
  iii. Оставшиеся 60 символов второго предложения.

- *Язык* — язык документов-источников. Настройка помогает разбить текст на чанки корректно. Если ваши источники на нескольких языках, выберите язык, на котором будут чаще всего отправлять запросы базе знаний.

Подготовка данных с помощью LLM

> ⚠ **ПРЕДУПРЕЖДЕНИЕ**
> Включение опций может привести к значительному увеличению расходов.

- *Обогащать чанки* — добавлять к чанку дополнительную информацию для улучшения качества поиска: заголовок, краткое содержание, ключевые слова и вопросы, на которые отвечает чанк.
- *Генерировать описания изображений* — добавлять чанки с описанием изображений для поиска по ним. Подробнее в разделе Изображения в ответе.

Настройки LLM

Настройки LLM будут применяться:

- для формирования чанков, если выбран способ разделения текста с помощью LLM;
- для обогащения чанков, вне зависимости от выбранного способа разделения.

> **⚠ ПРЕДУПРЕЖДЕНИЕ**
>
> Эти настройки не применяются для генерации описаний изображений.

Состав настроек:

- *Модель* — выберите одну из доступных языковых моделей.
- *Максимальное количество токенов в запросе* — ограничивает количество токенов, которое может быть отправлено в LLM.
- *Максимальное количество токенов в ответе* — ограничивает количество токенов, которое может быть сгенерировано LLM за одну итерацию.
- *Температура* — регулирует креативность ответов. При более высоких значениях результаты будут более творческими и менее предсказуемыми.

> **ПОДСКАЗКА**
>
> Чтобы посмотреть, на какие части разделен ваш источник, скачайте архив с чанками:
>
> 1. Перейдите в раздел *Источники* и наведите курсор на нужный источник.
> 2. Нажмите → *Архив чанков*.
>
> При тестировании базы знаний вы также сможете посмотреть, какие чанки были выбраны для подготовки ответа.

Настройки поиска чанков

Чтобы перейти к параметрам поиска чанков, выберите в меню *Настройки проекта → Поиск*.

Выбор пайплайна определяет последовательность действий, выполняемых для извлечения чанков и подготовки ответа:

- *Семантический пайплайн* — система сравнивает векторное представление (эмбеддинг) запроса пользователя с векторными представлениями чанков. Наиболее релевантные чанки вместе с запросом будут переданы в LLM для генерации ответа.
- *Агентский пайплайн* — AI-агент управляет процессом подготовки ответа, используя LLM и вызовы функций. LLM через AI-агента инициирует поиск чанков, анализирует их и запрашивает дополнительные. Когда LLM считает полученные чанки наиболее подходящими, подбор чанков завершается, и LLM генерирует ответ.

Более подробное описание пайплайнов см. в разделе Основные этапы работы Jay Knowledge Hub.

Состав настроек зависит от выбранного пайплайна.

Семантический пайплайн **Агентский пайплайн**

Параметры поиска чанков в источниках
Повторное ранжирование
Аббревиатуры

Параметры поиска чанков в источниках

- *Количество результатов поиска* — количество найденных чанков, которые будут переданы для дальнейшей обработки.

  > **ПОДСКАЗКА**
  >
  > Если вы планируете использовать повторное ранжирование, увеличьте количество результатов поиска. Это позволит выбирать подходящие чанки из большего числа кандидатов.

- *Количество соседних чанков* — количество соседних чанков каждого найденного чанка, которые будут переданы вместе с ним. Соседние чанки помогают модели лучше понять контекст запроса и сгенерировать более точный ответ.

Например, если *Количество результатов поиска* равно 10, а *Количество соседних чанков* — 3, то к каждому найденному чанку добавится еще по 3 чанка до и после него. Таким образом объем найденного чанка увеличится, в итоге для дальнейшей обработки будут переданы 10 «больших» чанков.

- *Поиск по метаданным чанков* — если включен, система будет искать подходящие чанки не только по их контенту, но и по дополнительным полям, которые были заполнены в процессе индексации.
- *Поиск по сегментам* — определяет поведение базы знаний в случае, если не найдены чанки в заданной области поиска:
  - *Расширяемый* — поиск продолжится по всей базе знаний.
  - *Строгий* — база знаний ответит, что данных нет.

  Ограничить область поиска чанков можно в запросе к базе знаний через API и в тестовом диалоге. В область поиска можно включить как определенные сегменты, так и источники, у которых сегменты не заданы. Подробнее читайте в разделе Сегменты.

  > **ПОДСКАЗКА**
  >
  > Для расширяемого режима рекомендуем включить повторное ранжирование. Если чанки, найденные в заданной области поиска, недостаточно релевантны запросу, повторное ранжирование отфильтрует их и сработает поиск по всей базе знаний.

- *Перефразировать запрос пользователя* — если включено, система попытается перефразировать запрос пользователя, чтобы он стал понятнее. Это может улучшить качество поиска.

  Вы можете отредактировать промт для перефразирования запроса, если требуется.
- *Полнотекстовый поиск* — если включен, в дополнение к семантическому поиску будет использоваться полнотекстовый.

  Семантический поиск отбирает чанки, подходящие по смыслу, даже если слова из запроса отсутствуют. Полнотекстовый поиск учитывает только совпадения со словами из запроса.

  Выберите стратегию использования полнотекстового поиска:

  - *Гибридный* поиск объединяет результаты семантического и полнотекстового поиска, предоставляя как чанки, подходящие по смыслу, так и содержащие слова из запроса. Итоговое количество результатов не превышает *Количество результатов поиска*, указанное выше.
  - *Взвешенный* поиск позволяет комбинировать результаты семантического и полнотекстового поиска в нужном соотношении.

    Укажите максимальное количество результатов каждого вида поиска, которое можно включить в итоговый список для дальнейшей обработки.

Например, если *Количество результатов поиска* равно 10, *Порция семантики* — 8, а *Порция полнотекстового поиска* — 4, вначале отбирается по 10 результатов каждого вида поиска. Затем результаты объединяются, отбрасываются повторы (чанки, найденные обоими видами поиска) и результаты сверх указанных порций. Итоговое количество результатов — не более 12 (8 + 4).

- *С порогом*: в первую очередь выполняется семантический поиск, но если он не предоставил достаточно релевантных результатов — применяется полнотекстовый.

  Укажите пороговую оценку релевантности для результатов семантического поиска. Если все результаты ниже порога, поиск переключается на полнотекстовый.

- *Учет истории диалога* — если включен, система будет учитывать предыдущие запросы пользователя при поиске чанков. Это позволяет модели лучше понимать контекст и генерировать более точные ответы.

  Для учета истории вы можете модифицировать следующие настройки:

  - *Промт*;
  - *Максимальный размер истории диалога, в токенах*, который может быть отправлен в LLM;
  - *Минимальное количество запросов пользователя* в истории диалога.

  > **⚠ ПРЕДУПРЕЖДЕНИЕ**
  >
  > Если размер истории превышает ограничение по токенам, сообщения удаляются по одному, пока не будет достигнут нужный размер или минимальное количество запросов пользователя. Если минимальное количество запросов достигнуто раньше, история отправляется в LLM с превышением размера, что может привести к ошибке. Если ошибка повторяется, рекомендуем уменьшить минимальное количество запросов в истории.

Повторное ранжирование

Повторное ранжирование выполняется после извлечения чанков из документов-источников, чтобы заново оценить их релевантность запросу пользователя и отобрать наиболее подходящие. Если найденные чанки соответствуют запросу достаточно хорошо, повторное ранжирование можно отключить.

Выберите способ повторного ранжирования:

- *Эмпирическое* — используется алгоритм, основанный на опыте Just AI.
- *С помощью модели* — используется специальная модель-реранкер.

Эмпирическое ранжирование работает быстрее, чем ранжирование с помощью модели, но менее точно.

Настройки повторного ранжирования зависят от выбранного способа.

Эмпирическое **С помощью модели**

- *Количество результатов поиска* — максимальное количество результатов, которые после повторного ранжирования будут отправлены в LLM вместе с запросом пользователя для генерации ответа.
- *Количество результатов из одного документа* — настройка полезна, чтобы чанки из одного документа-источника не заняли весь контекст при подготовке ответа.
- *Пороговая оценка* — результаты с оценкой релевантности ниже порога не будут использоваться для генерации ответа.
- *Максимальное отличие от лучшей оценки, в %* — допустимое отклонение от оценки самого релевантного чанка. Результаты, оценка которых отличается сильнее, не будут использоваться для генерации ответа.

Аббревиатуры

Если в ваших документах-источниках используются аббревиатуры и другие сокращения, которые отличаются от общепринятых, вы можете сделать их понятными системе. Для этого загрузите список таких аббревиатур с расшифровкой в виде файла формата JSON. Шаблон файла можно скачать в интерфейсе.

▼ Формат файла

Каждый JSON-объект описывает аббревиатуру. Ключ объекта — сама аббревиатура.

Значение содержит поля:

- `name` — расшифровка аббревиатуры.
- `variations` — варианты написания.
- `description` — объяснение понятия.

Если в запросе пользователя найдена аббревиатура или вариант из `variations`, они добавляются в промт для перефразирования вместе с `name` и `description`.

Пример файла:

```
{
  "ИНН": {
```

```
    "name": "Идентификационный номер налогоплательщика",
    "variations": [],
    "description": "Налоговый номер"
  },
  "PIN": {
    "name": "PIN-код",
    "variations": ["пин"],
    "description": "Цифровой пароль для банковской карты"
  }
}
```

Настройки генерации ответов

Чтобы перейти к параметрам генерации ответов на запросы пользователей, выберите в меню *Настройки проекта → Генерация*.

Системный промт

Если требуется, вы можете отредактировать системный промт, который LLM будет использовать при генерации ответа на запрос пользователя.

Настройки LLM

Настройки LLM будут применяться:

- для генерации ответа на запрос пользователя;
- для поиска чанков, если выбран агентский пайплайн;
- для перефразирования запроса и учета истории, если выбран семантический пайплайн.

Основные настройки:

- *Модель* — выберите одну из доступных языковых моделей. Для агентского пайплайна доступны только модели, которые поддерживают function calling — он позволяет модели запрашивать чанки.
- *Максимальное количество токенов в запросе* — ограничивает количество токенов, которое может быть отправлено в LLM.
- *Максимальное количество токенов в ответе* — ограничивает количество токенов, которое может быть сгенерировано LLM за одну итерацию.
- *Температура* — регулирует креативность ответов. При более высоких значениях результаты будут более творческими и менее предсказуемыми. Рекомендуем менять или температуру, или Top P, но не оба параметра одновременно.

Расширенные настройки:

- *Top P* — регулирует разнообразность ответов. При низких значениях нейросеть выбирает из меньшего количества вероятных слов, но при высоких — ответ может получиться разнообразнее. Рекомендуем менять или Top P, или температуру, но не оба параметра одновременно.

- *Presence penalty* — штраф за повторное появление токенов. Чем выше значение, тем меньше вероятность повторения слов или фраз в ответе.

  Все повторы штрафуются одинаково, независимо от частоты. Например, появление токена во второй раз будет оштрафовано так же, как в 10-й.
- *Frequency penalty* — штраф за частоту появления токенов. Чем выше значение, тем меньше вероятность появления слов или фраз, которые уже встречались в ответе несколько раз.

  Влияние Frequency penalty возрастает в зависимости от того, сколько раз токен уже появился в тексте.

Показывать документы-источники в ответе бота

Если включено, ответ базы знаний будет дополнен списком источников — файлов или страниц, на основе которых он сгенерирован.

<details>
<summary>Как представлены источники в ответе</summary>

Список источников содержит их названия и ссылки. В API Jay Knowledge Hub список источников возвращается в виде массива `relevantSources`.

- Если источник загружен из интеграции, выдается ссылка на оригинал, например на страницу или вложенный файл в Confluence.
- Если источник загружен вручную в виде файла, в каналах и в API выдается временная ссылка для скачивания. Срок действия таких ссылок ограничен. В тестовом диалоге отображается ссылка на раздел *Источники* и кнопка для скачивания.

Дополнительно в API доступен метод для выгрузки источника из базы знаний: GET /sources/{sourceId}/download.
</details>

Сегменты базы знаний

Сегмент — это тема или категория, к которой относится документ-источник в базе знаний. Используйте сегменты, чтобы выделить в базе знаний логические разделы и разграничить доступ к информации.

Источник можно включить в один или несколько сегментов или оставить без сегмента.

Примеры сегментов:

- По целевой аудитории: «Физические лица», «Юридические лица», «Партнеры», «Технические специалисты».
- По продуктам и услугам: «Кредитование», «Инвестиции».
- По процессам: «Закупки», «Контроль качества».

Формирование сегментов

Укажите названия сегментов при загрузке файлов и настройке интеграций. Предварительно регистрировать сами сегменты не требуется, названия могут быть произвольными.

Чтобы изменить сегменты после загрузки источников:

1. Перейдите в раздел *Источники* в проекте базы знаний.
2. Отметьте источники и нажмите *Сегменты*.
3. Задайте новый набор сегментов. Прежний набор сегментов будет удален и заменен новым.

При изменении набора сегментов в настройках интеграции новый набор применяется ко всем источникам из этой интеграции — как уже загруженным, так и будущим.

Поиск по сегментам

Вы можете задать область поиска в запросе к базе знаний через API и в тестовом диалоге.
В область поиска можно включить как определенные сегменты, так и источники, у которых сегменты не заданы.

В настройках проекта доступно два режима поиска:

- *Строгий* — поиск выполняется только в заданной области. Если в ней не найдено релевантной информации, система сообщает об отсутствии данных.
- *Расширяемый* (выбран по умолчанию) — если в заданной области не найдено ни одного чанка, поиск выполняется по всей базе знаний.

  > ⓘ **ОСОБЕННОСТИ РАСШИРЯЕМОГО ПОИСКА**
  >
  > Система может найти в заданной области чанки с низкой релевантностью. Поскольку хотя бы один чанк найден, поиск не будет расширен на всю базу знаний. В результате ответ может быть неполным или неточным.
  >
  > Чтобы повысить качество поиска в таких случаях, включите повторное ранжирование. Он оценит релевантность найденных чанков. Если реранкер отфильтрует их как неподходящие, система запустит поиск по всей базе знаний.

Оценка качества

Чтобы оценить качество ответов по отдельным сегментам базы знаний, добавьте в тест-сет столбец *Segments*. Если для ответа на вопрос нужно искать информацию в определенных сегментах, укажите их через запятую. Для поиска по источникам, которые не относятся ни к одному сегменту, укажите `include_without_segments`. Для поиска по всей базе знаний оставьте поле пустым.

Пример значения: `Контроль качества,include_without_segments`.

© ООО «Маинд Крафт», 2025