Описание endpoint поиска в semantic-search-service

Все запросы должны выполняться с header Authorization со значением “Bearer <access токен, полученный после создания интеграции>“. Если параметр опциональный, то при его отсутствии будет использоваться значение по умолчанию.

POST /api/v1/semantic/external/search

Параметры запроса:

|  |  |  |  |
| --- | --- | --- | --- |
| **Параметр** | **Тип данных** | **Описание** | **Опционально** |
| query | string | Поисковый запрос | - |
| limit\_count | integer | Лимит ответа по количеству чанков\* (по умолчанию 100) | + |
| limit\_tokens | integer | Лимит ответа в токенах (по умолчанию 0, не ограничевается). | + |
| user\_id | uuid | ID пользователя, у которого будут проверяться права на статьи (по умолчанию берётся ID пользователя, от которого была создана внешняя интеграция) | + |
| check\_permissions | bool | Проверять ли доступ к статьям относительно пользователя с user\_id? (По умолчанию false) | + |

\*Чанк — это часть статьи, которая была проиндексирована. Индексация статей происходит не целиком, а чанками, делящими тело статьи на абзацы и предложения. Если предложение больше 512 токенов, оно делится, и оставшаяся часть переходит в следующий чанк. Максимальный размер чанка составляет 512 токенов. Для повышения качества поиска используется „перехлест“ чанков, это позволяет добиться большей семантической близости чанков одного документа.

Примечание: При limit\_tokens = 0 запрос не обрезается по количеству токенов.

Если указан limit\_tokens > 0, то из результатов, полученных с ограничением по limit\_count, будут возвращены ответы в порядке уменьшения score, которые по сумме токенов проходят в указанный limit\_tokens.  
Если сразу же chunk\_token\_length первого релевантного чанка будет больше указанного limit\_tokens, то вернётся пустой ответ.

Пример:

* limit\_count = 50
* limit\_tokens = 512

Из 50 наиболее релевантных результатов в порядке уменьшения score будет проверяться ограничение на limit\_tokens, чтобы сумма chunk\_token\_length не превышала 512. Если у первого релевантного чанка в массиве будет chunk\_token\_length 513 токенов, то вернётся пустой массив.

Пример запроса:

{

"query": "ии в медицине",

"limit": 500,

"user\_id": "b3c18437-5765-4f55-bd5d-2f54340c45bc",

"check\_permissions": false

}

Содержание ответа:

|  |  |  |
| --- | --- | --- |
| **Параметр** | **Тип данных** | **Описание** |
| container\_id | uuid | ID контейнера, в котором находится источник (пространство для статей, диск для документов, ID самой ТБД для ТБД) |
| source\_id | uuid | ID источника |
| title | string | Название источника |
| score | number | Скорринг чанка источника до реранкинга |
| type | string | Тип источника (статья - article - документ, doc, ТБД - content-database) |
| rerank\_score | number | Скорринг чанка источника после реранкинга (если использование реранкера отключено, то null) |
| final\_score | number | Итоговый скорринг чанка (если использование реранкера отключено, то null) |
| text | string | Текст чанка статьи |
| chunk\_token\_length | integer | Длина чанка статьи в токенах |
| offset | integer | Количество символов от начала статьи |
| length | integer | Длина ответа в символах (без спецсимволов) |
| link | string | Путь до источника |
| properties | string | Свойства сущности (передаются только для статей из ТБД) |

Пример ответа:

[

{

"container\_id": "0891797d-536a-411e-8954-f58d957e3784",

"source\_id": "eb211763-07d7-4bbb-bc14-d1d6df791f98",

"title": "Искусственный интеллект.docx",

"score": 2.178287,

"type": "doc",

"rerank\_score": 2.623046875,

"final\_score": 5.801333875,

"text": "Искусственный интеллект (ИИ) — это область компьютерных наук, которая занимается созданием интеллектуальных машин, способных выполнять задачи, обычно требующие человеческого интеллекта. ИИ включает в себя множество подполей, таких как машинное обучение, обработка естественного языка, компьютерное зрение, робототехника и многие другие. Машинное обучение, одно из ключевых направлений ИИ, сосредоточено на создании алгоритмов, которые позволяют компьютерам обучаться на данных и делать прогнозы или принимать решения без явного программирования. Это включает в себя методы обучения с учителем, без учителя и с подкреплением.\nОбработка естественного языка (NLP) — это область ИИ, которая занимается взаимодействием между компьютерами и человеческим языком. NLP включает в себя задачи, такие как перевод текста, анализ тональности, распознавание речи и генерация текста. Компьютерное зрение — это область ИИ, которая занимается обработкой и анализом визуальных данных, таких как изображения и видео. Она включает в себя задачи, такие как распознавание объектов, сегментация изображений и распознавание лиц.\n| | | | | | | |\n| --- | --- | --- | --- | --- | --- | --- |\n| | | | | | | |\n| | | | | | | |\n| | | | | | | |\n\n\nРобототехника — это область ИИ, которая занимается созданием роботов, способных выполнять задачи автономно или полуавтономно. Роботы могут использоваться в различных областях, таких как производство, медицина, исследование космоса и бытовые услуги. ИИ находит применение в различных отраслях, таких как здравоохранение, финансы, транспорт и развлечения. В здравоохранении ИИ используется для диагностики заболеваний, разработки лекарств и персонализированной медицины.\nВ финансах ИИ используется для анализа рынков, управления рисками и обнаружения мошенничества. В транспорте ИИ используется для автономного вождения, управления трафиком и логистики. В развлечениях ИИ используется для создания персонализированных рекомендаций, генерации контента и улучшения пользовательского опыта. ИИ также поднимает этические и социальные вопросы, такие как конфиденциальность данных, безопасность и влияние на рынок труда.\nВажно разрабатывать и использовать ИИ ответственно, учитывая этические и социальные последствия. Будущее ИИ включает в себя развитие более сложных и мощных алгоритмов, интеграцию ИИ в различные аспекты повседневной жизни и решение глобальных проблем, таких как изменение климата и бедность. ИИ также будет играть ключевую роль в развитии новых технологий, таких как квантовые вычисления и интернет",

"chunk\_token\_length": 614,

"offset": 0,

"length": 2553,

"link": "/space/02cb7763-12ae-4bf4-9ee6-aadf59756a0a/article/b82593a9-465f-46b9-b906-e53547d96f02#ac32a2c1-0ac3-4d70-a301-24bc583b8f69",

"properties": ""

}

]