Настройки знаний

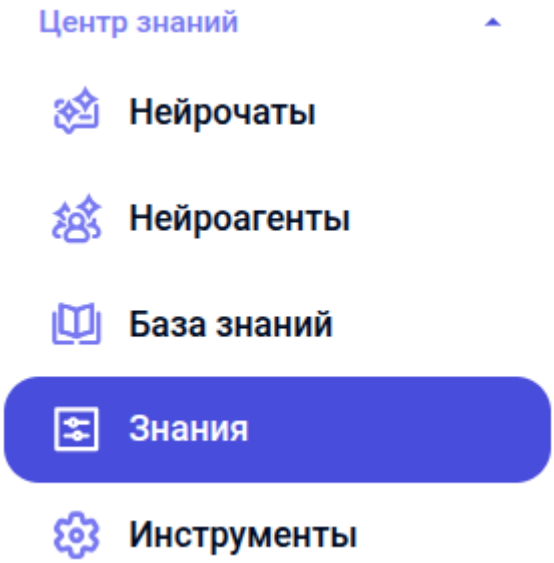
> **Общая информация о «Знаниях»**
>
> Раздел **«Знания»** в Центре знаний позволяет управлять тем, как нейроагенты находят информацию в подключённых документах и используют её при ответах пользователям.
> Каждое знание содержит:
>
> - список документов для поиска;
> - параметры работы RAG-поиска;
> - настройки формирования контекста и результата.

Настройки можно изменить **на странице настройки знания** или **во время тестирования знания**.
Интерфейсы различаются:

- **настройка знания** — для конфигурации параметров;
- **тестирование знания** — для проверки качества поиска и корректности выдачи.

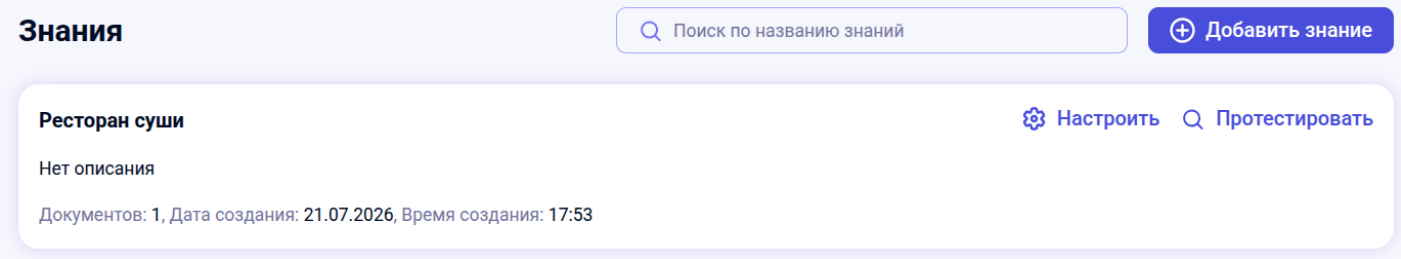
---

1. Как открыть настройки знания



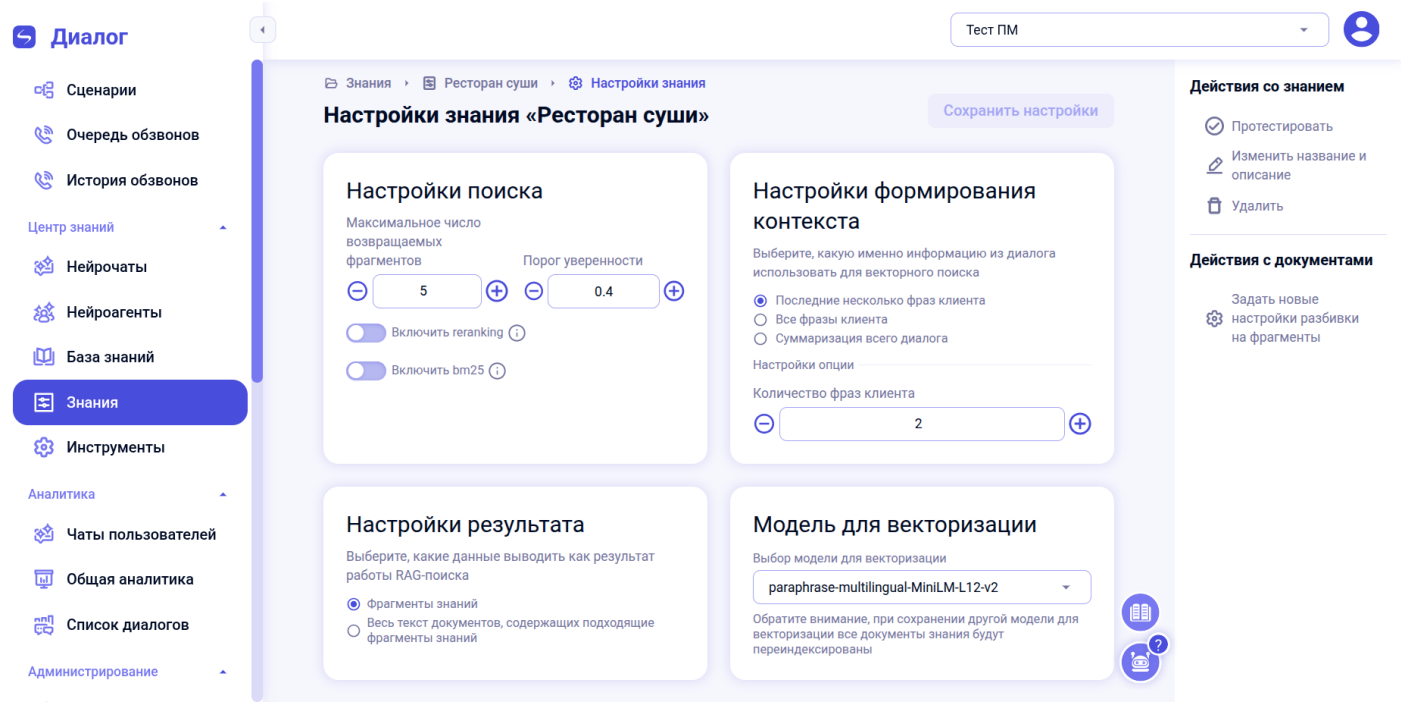
1. Перейдите в раздел **Центр знаний → Знания**.

2. Выберите нужное знание из списка.
3. В правом верхнем углу нажмите **«Изменить настройки»** или откройте вкладку **«Настройки знания»**.



Вы попадёте на страницу с четырьмя блоками:

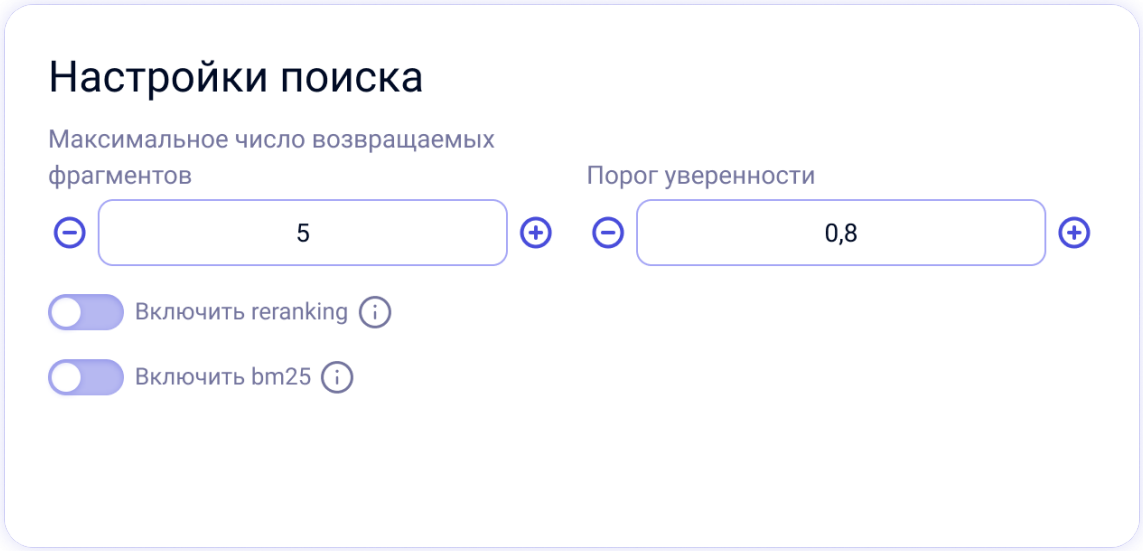
- **Настройки поиска**
- **Настройки формирования контекста**
- **Настройки результата**
- **Модель для векторизации**



После внесения изменений нажмите **«Сохранить настройки»**.

---

2. Настройки поиска



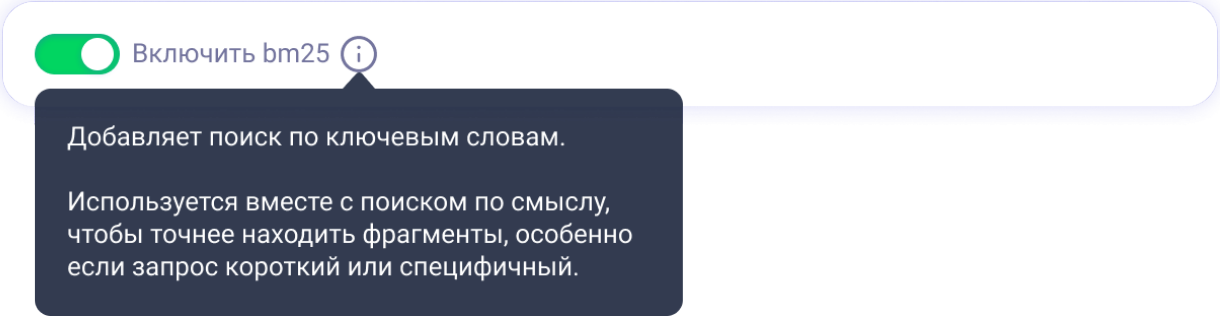
Этот блок определяет, **как именно система ищет фрагменты знаний** в подключённых документах.

Основные параметры

- **Максимальное число возвращаемых фрагментов знаний**
  Определяет, сколько фрагментов знаний система вернёт по каждому запросу.
  Обычно 3–5 достаточно для качественного ответа.
- **Порог уверенности (0–1)**
  Минимальный уровень смыслового совпадения.
  Фрагменты с уверенностью ниже указанного значения не попадают в результат.
  Например, при 0,8 будут использованы только наиболее релевантные фрагменты.

Дополнительные параметры

BM25 — поиск по ключевым словам



Включите, если хотите дополнить поиск по смыслу поиском по формулировкам.

- BM25 ищет совпадения по словам и используется **вместе** с векторным поиском.
- Повышает точность при коротких или специфичных запросах.

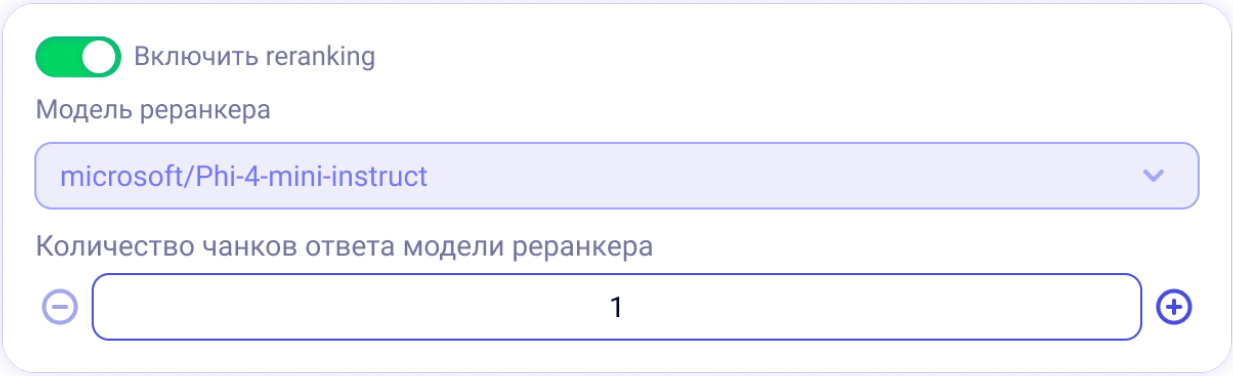
Плюсы:

- Находит редкие слова и точные формулировки.
- Эффективен, если запрос пользователя короткий.

Минусы:

- Увеличивает время ответа.

Reranking — повторная сортировка результатов



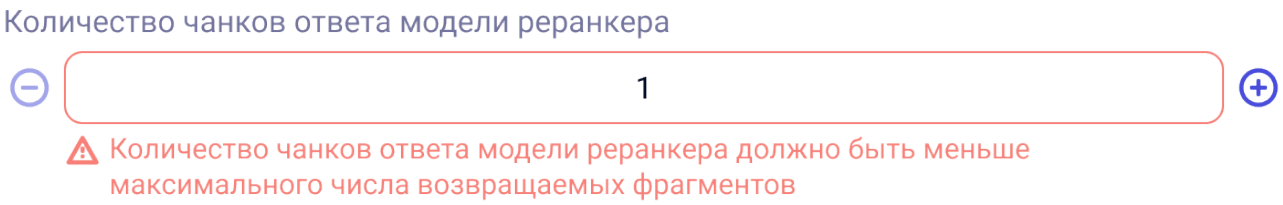
Опция улучшает качество поиска за счёт повторной проверки уже найденных фрагментов знаний.

При включении:

- система использует отдельную модель реранкера (например, *microsoft/Phi-4-mini-instruct*);
- найденные фрагменты проверяются повторно и сортируются по смысловой близости.

Дополнительные параметры:

- **Модель реранкера** — выбирается из списка.
- **Количество фрагментов знаний для реранкера** — должно быть меньше общего числа, указанного в параметре выше.
  Если задано большее — появится предупреждение, и система не позволит сохранить настройки.



Плюсы:

- Повышает точность выдачи.
- Снижает риск появления нерелевантных фрагментов.

Минусы:

- Увеличивает время ответа (до 1 секунды).

---

3. Настройки формирования контекста

Определяют, какую часть пользовательского диалога система использует для поиска по знаниям.

Доступны три режима:

- **Последние несколько фраз клиента** — используется ограниченное число последних сообщений. Укажите количество в поле «Количество фраз клиента». Подходит для коротких диалогов, где контекст быстро меняется.
- **Все фразы клиента** — учитывается весь диалог. Используется в длинных или тематически связанных обращениях.
- **Суммаризация всего диалога** — весь диалог сжимается в краткую сводку отдельной моделью (поле «Модель суммаризации»), и для поиска используется эта сводка, а не исходные фразы целиком. Рядом — ссылка «Расширенные настройки модели» (тот же набор параметров, что и у блока «Диалог с LLM»: temperature, top_p, max_new_tokens и др.). Подходит для длинных диалогов, где важен смысл целиком, но не нужно перегружать поиск объёмом текста.

Рекомендации:

- Для справочных чат-ботов — 2–3 последних фразы.
- Для сложных консультаций — весь диалог или суммаризация, если диалог длинный.

---

4. Настройки результата

Определяют, **в каком виде найденные данные передаются нейроагенту**.

- **Фрагменты знаний** — система возвращает только найденные текстовые части.
  Это быстрый и экономичный вариант, подходящий для коротких ответов.
- **Весь текст документов, содержащих подходящие фрагменты знаний** —
  возвращается полный текст документов, где найдены совпадения.
  Этот режим реализует принцип **Parent-child chunking**:
  - *child* — найденный фрагмент знаний;
  - *parent* — документ, в котором он находится.
    Система передаёт нейромодели весь родительский документ, чтобы сохранить контекст.

Плюсы:

- Помогает избежать потери смысла при сложных вопросах.
- Повышает полноту и корректность ответа.

Минусы:

- Увеличивает объём передаваемых данных.
- Может замедлить время генерации ответа.

---

5. Модель для векторизации

Определяет, какая модель используется для преобразования текста документов в векторное представление — то есть **поиск по смыслу**.

- Выберите нужную модель в выпадающем списке.
- При изменении модели необходимо подтвердить **реиндексацию** — система пересчитает векторы всех документов знания.

Во время пересчёта знание временно недоступно для использования.

---

6. Тестирование знания

После сохранения настроек можно проверить, как работает поиск.

1. На странице тестирования знания введите запрос в строку поиска.
2. Система покажет список найденных документов и фрагментов знаний.
3. Результаты позволяют оценить, насколько хорошо работает текущая конфигурация.

Если результаты не удовлетворяют:

- уменьшите порог уверенности;
- увеличьте количество возвращаемых фрагментов знаний;
- включите BM25 или Reranking.

---

7. Применение и сохранение

После внесения изменений нажмите **«Сохранить настройки»** (на странице настройки) или **«Применить к знанию»** (на странице тестирования).

Появится окно подтверждения — после применения все боты и сценарии, использующие это знание, начнут работу с новыми параметрами.

□□ Сравнение методов

Метод	Назначение	Преимущества	Ограничения
BM25	Поиск по ключевым словам	Хорош при коротких или редких запросах	Увеличивает время ответа, неэффективен при длинных запросах
Reranking	Повторная сортировка найденных фрагментов знаний	Повышает точность и качество ответов	Увеличивает время отклика
Parent-child chunking	Возврат целого документа с найденным фрагментом	Сохраняет контекст, предотвращает искажения смысла	Увеличивает объём данных и время обработки

Версия #6
editor2 создал 2025-10-08 06:56:12 UTC
editor1 обновил 2026-09-04 11:55:53 UTC