Ай-ТИ Фреш

ТЕХНИЧЕСКИЙ РАЗБОР

RAG-чат-бот по регламентам компании: наша методика

От шаринг-диска к чат-боту: чанкинг, эмбеддинги, rerank и ACL на практике

Июль 2026

itfresh.ru · ИТ-аутсорсинг для юридических лиц

О ЧЁМ ЭТОТ РАЗБОР

Суть проблемы

Сотрудник тратит 10-20 минут, чтобы найти актуальную версию регламента, прайса или инструкции среди ревизий на шаринг-диске, а часто находит устаревший файл и делает по нему ошибку. Мы строим для клиента внутреннего RAG-чат-бота: он ищет ответ по корпусу документов компании и возвращает конкретный пункт регламента со ссылкой на источник, а не общими словами от модели.

Почему это важно бизнесу

- Новый сотрудник неделями ищет, где регламент по закупкам, а не в чём его суть
- Разные отделы дают клиенту разные ответы по одному и тому же прайсу
- Устаревшие версии файлов на диске приводят к ошибкам в заказах и расчётах
- Держать штатного человека на актуализации wiki дороже, чем поддерживать индекс
- При проверках нельзя быстро показать, какой регламент действовал на дату события

ЧТО МЫ ЗАКЛАДЫВАЕМ

Ключевые параметры реализации

3-6 сек

среднее время ответа бота на вопрос по регламенту вместо 10-20 минут поиска

наши замеры, внедрения 2026

150-300

документов в типовом корпусе клиента до 50 рабочих мест на старте

наша практика аудита корпуса

~90%

вопросов сотрудников закрывается ботом без обращения к HR или бухгалтерии

статистика наших ботов, 2026

400-512

токенов — рабочий размер чанка регламента в нашем пайплайне

наша методика чанкинга

<3%

доля промахов в топ-10 после reranker по нашим пилотным замерам

внутренние тесты retrieval

5-7 дней

типовой срок запуска MVP-бота на корпусе клиента до 50 PM

наш таймлайн внедрения

РЕАЛИЗАЦИЯ 1 ИЗ 3

Оптовая торговая компания, ~40 РМ — прайсы и регламенты снабжения

Что настраиваем

Оптовая торговая компания, Москва, отдел снабжения и логистики

Как мы это делаем

1. Собрали корпус с 3 шаринг-дисков: 220 файлов docx/xlsx/pdf, до трети — дубли и старые версии прайсов
2. Дедуплицировали по хэшу содержимого и дате правки, оставили по одной актуальной версии на документ
3. Разметили ACL в метаданных чанков (снабжение/бухгалтерия/склад) ещё на этапе индексации, до эмбеддинга
4. Собрали гибридный индекс: pgvector HNSW + BM25, contextual retrieval через Claude Haiku 4.5
5. Подняли Telegram-бота с Citations поверх Claude Sonnet 4.5, ответы со ссылкой на пункт документа

РЕЗУЛЬТАТ

За 5 недель охватили весь корпус снабжения; повторные вопросы менеджеров по актуальности прайса в общий чат отдела практически исчезли, часть обращений к закупщику ушла в бота

КЛЮЧЕВОЙ НЮАНС

Главный риск оказался не в модели, а в качестве исходников: без чистки дублей и старых версий бот путал устаревший прайс с текущим

РЕАЛИЗАЦИЯ 2 ИЗ 3

Производственная компания, 45 РМ — регламенты охраны труда

Что настраиваем

Производственная компания, участок с посменной работой, Московская область

Как мы это делаем

1. Прогнали архив PDF-сканов старых инструкций через OCR перед индексацией — треть корпуса иначе была бы невидима для эмбеддингов
2. Ввели ABAC-метаданные (цех, должность, уровень допуска), фильтр применяется до ANN-поиска, а не после
3. Настроили чанкинг 512 токенов с перекрытием 80 токенов и contextual retrieval с заголовком цеха и раздела
4. Reranker voyage rerank-2.5-lite отсеивал отменённые редакции инструкций, оставшиеся в архиве по ошибке

РЕЗУЛЬТАТ

Бот стал первой точкой ответа по технике безопасности на участке, мастера перестали звонить в отдел охраны труда по типовым вопросам смены

КЛЮЧЕВОЙ НЮАНС

Без этапа OCR треть регламентов в виде сканов осталась бы вне поиска — проверку «бот не находит документ X» мы теперь делаем обязательным шагом приёмки

РЕАЛИЗАЦИЯ 3 ИЗ 3

IT-компания на аутсорсе, 28 PM — кадровые регламенты и инструкции

Что настраиваем

Компания-заказчик ИТ-аутсорсинга, HR-департамент

Как мы это делаем

1. Настроили отслеживание правок на диске (диф по хэшу файла раз в час) вместо ручной пересборки индекса
2. Изменённые документы переиндексируются точечно — только затронутые чанки, а не весь корпус
3. Включили prompt caching для системного промпта и часто цитируемых регламентов — заметно снизили счёт за токены
4. Добавили аудит-лог: user_id, хэш запроса, id выданных и отфильтрованных по ACL чанков

РЕЗУЛЬТАТ

HR перестал вручную рассылать актуальные версии положений при каждом изменении — бот всегда отвечает по последней редакции

КЛЮЧЕВОЙ НЮАНС

Инкрементальная переиндексация оказалась важнее выбора модели — без неё корпус в 300+ файлов пересчитывался бы часами при каждой правке

ЧТО МЫ ВЫУЧИЛИ НА ПРАКТИКЕ

Подводные камни

- ✕ **Чанкинг без учёта структуры документа**
  Резать регламент по фиксированной длине символов без учёта таблиц и пунктов — фраза или строка таблицы рвётся пополам
- ✕ **Фильтр доступа после поиска, а не до**
  ACL-проверка постфактум на выдаче — модель уже увидела чужой документ до фильтрации, риск утечки через ответ
- ✕ **Один эмбеддинг без reranker**
  Топ-20 по косинусной близости без переранжирования даёт заметный процент нерелевантных чанков в контексте
- ✕ **Нет OCR для сканов и фото инструкций**
  Часть регламентов — сканы, PDF без текстового слоя; без OCR они просто не попадают в индекс
- ✕ **Ручная пересборка индекса при правках**
  Обновление регламента вручную запускает переиндексацию всего корпуса вместо точечного апдейта изменённых чанков
- ✕ **Дубли и устаревшие версии в корпусе**
  Индексируем всё подряд с диска — бот случайно цитирует прайс годовой давности наравне с актуальным
- ✕ **Промпт-кэш не настроен**
  Системный промпт и часто цитируемые документы гоняются заново на каждый запрос — счёт за токены растёт кратно
- ✕ **Нет аудита ответов и источников**
  Без Citations и лога выданных чанков невозможно проверить, откуда бот взял факт, когда он ошибся

УРОВНИ РЕАЛИЗАЦИИ

Как правильно

МИНИМУМ

- Собрать и дедуплицировать корпус документов, убрать устаревшие версии
- Chunking 400-512 токенов с overlap 10-20%, recursive splitter по структуре
- Гибридный поиск: эмбеддинги + BM25 без reranker на пилоте

НОРМАЛЬНО

- Добавить reranker (voyage rerank-2.5-lite) перед подачей в LLM
- Contextual retrieval: LLM-контекст 50-100 токенов на чанк перед эмбеддингом
- ACL-метаданные на чанк, фильтр до ANN-поиска, не после
- Citations API для ссылок на источник в каждом ответе

ХОРОШО

- Инкрементальная переиндексация по диффу хэшей вместо полной пересборки
- Prompt caching системного промпта и топ-документов для снижения счёта
- Аудит-лог запросов, выданных и отфильтрованных чанков по ACL
- OCR-конвейер для сканов + регулярная проверка «слепых зон» корпуса

ПРОВЕРЬТЕ У СЕБЯ

Чек-лист самопроверки

- [ ] Провести аудит корпуса: сколько файлов, сколько дублей, сколько сканов без текстового слоя
- [ ] Прогнать OCR по сканам перед индексацией, иначе часть регламентов не попадёт в поиск
- [ ] Разметить ACL/классификацию документов на этапе индексации, а не постфактум
- [ ] Настроить recursive chunking 400-512 токенов, overlap 10-20%
- [ ] Включить contextual retrieval — короткий заголовок раздела и документа в каждом чанке
- [ ] Поднять гибридный индекс (pgvector HNSW или Qdrant) + BM25
- [ ] Добавить reranker перед финальной подачей контекста в модель
- [ ] Включить Citations API, чтобы ответ ссылался на конкретный пункт документа
- [ ] Настроить инкрементальную переиндексацию по диффу изменений на диске
- [ ] Включить аудит-лог запросов и выданных чанков для проверки утечек и галлюцинаций

Если хотя бы на два вопроса ответ «нет» или «не знаю» — тема требует внимания.

ЧТО МЫ ДЕЛАЕМ

Как поможет ITFresh

ITFresh — ИТ-аутсорсинг для юридических лиц до 50 рабочих мест в Москве и области. 15+ лет практики, собственная инфраструктура в дата-центре МТС (8 серверов Dell Xeon Platinum).

- Аудит и очистка корпуса документов клиента перед запуском RAG-бота
- Настройка ACL-фильтрации по отделам и уровням доступа на этапе индексации
- Подбор и настройка стека: pgvector/Qdrant, contextual retrieval, reranker
- Интеграция бота в Telegram, внутренний портал или почту сотрудников
- Сопровождение: мониторинг качества ответов, переиндексация, аудит-логи

15+
лет в ИТ-поддержке

50
рабочих мест — наш профиль

МТС
дата-центр, Москва

КОНТАКТЫ

Обсудить вашу задачу

Сайт	**itfresh.ru**
Телефон	**+7 903 729-62-41**
Telegram	**@ITfresh_Boss**

Бесплатно посмотрим вашу инфраструктуру по этому чек-листу и скажем, где тонко — без обязательств.

itfresh.ru

ДОКУМЕНТАЦИЯ И НАШ СТЕК

Техническая база

01 Contextual Retrieval (anthropic.com/engineering — 2024-2026)

02 Models overview (platform.claude.com — 2026)

03 Citations (platform.claude.com — 2026)

04 Prompt caching (platform.claude.com — 2026)

05 Rerankers (rerank-2.5) (docs.voyageai.com — 2025-2026)

06 voyage-context-3 (docs.voyageai.com — 2025)

07 pgvector README (HNSW/IVFFlat) (github.com/pgvector — 2026)

08 Collections / Indexing (qdrant.tech/documentation — 2026)

Основано на официальной документации продуктов и нашей практике внедрения.