

Чек-лист готовности данных к RAG

Пятнадцать проверок, которые мы прогоняем до того, как согласиться делать поиск по документам. Отметьте то, что у вас уже есть, и посмотрите итог на второй странице.

Корпус: что вообще есть

- ☐ **01. Вы можете назвать точное место, где лежат документы, по которым нужны ответы.**
«Где-то на диске и в почте у Марины» — это не корпус.
- ☐ **02. Вы знаете, сколько их: сотни, тысячи или десятки тысяч.**
От порядка величины зависит нарезка и стратегия поиска.
- ☐ **03. Известно, какая доля — сканы и фотографии, а не текст.**
Скан — отдельный этап распознавания со своей стоимостью и ошибками.
- ☐ **04. Вы можете сказать, какая версия документа действующая.**
Отменённую редакцию система процитирует так же уверенно, как действующую.

Структура: как документ устроен

- ☐ **05. В документах есть заголовки разделов, а не сплошной текст.**
Заголовки — самая дешёвая граница для нарезки.
- ☐ **06. Вы знаете, есть ли таблицы и насколько они важны для ответов.**
Таблица, разрезанная пополам, даёт уверенно неверный ответ.
- ☐ **07. Понятно, есть ли перекрёстные ссылки между документами.**
«В порядке, установленном приложением 3» — без приложения ответ бесполезен.
- ☐ **08. Дубли и почти-дубли известны хотя бы приблизительно.**
Пять версий одного регламента вытесняют из выдачи всё остальное.

Права: кто что видит

- ☐ **09. Вы знаете, все ли документы можно показывать всем сотрудникам.**
Права нужно закладывать в поиск с первого дня, иначе переделывать индекс.
- ☐ **10. Известно, есть ли в корпусе персональные данные.**
Это решает, где корпус может обрабатываться и какую модель можно взять.
- ☐ **11. Есть человек, который отвечает за содержание документов.**
Без владельца контента спор о неверном ответе упирается в разработчика.
- ☐ **12. Понятно, как часто документы меняются.**
От этого зависит, нужен ли инкрементальный доиндекс.

Проверяемость: как узнать, что получилось

- ☐ **13. Вы можете записать двадцать реальных вопросов, которые задают люди.**
Не выдуманных, а тех, что приходили в поддержку. Это и есть ТЗ.
- ☐ **14. Для каждого вопроса известен документ с правильным ответом.**
Это превращает «нравится или нет» в измеримое.
- ☐ **15. Есть договорённость, что считается приемлемым ответом.**
Критерий: нужный документ в первых трёх, ответ со ссылкой, честное «не нашёл».

Как читать результат

12 и больше «да»

Данные готовы. Начинайте с пилота на одном разделе корпуса и наборе из двадцати вопросов: за две-три недели видно, попадает ли поиск.

От 7 до 11

Обычный расклад для живой компании. Начинайте с той части корпуса, где ответов «да» больше всего, параллельно закрывая права и версию.

Меньше 7

RAG сейчас даст красивое демо и неработающую систему. Дешевле потратить две-три недели на порядок в документах: владелец, версии, действующие редакции. Это работа не про AI, и без неё результат не изменится.

Прошли чек-лист?

Напишите, что получилось: сколько документов, какие ответы «нет» и какие вопросы должна закрывать система. Ответим разбором по вашему корпусу — что делать в первую очередь и нужен ли вам вообще RAG.

ioann-ai.ru/instrumenty/chek-list-gotovnosti-k-rag