



команда роста на аутсорсе
для системного улучшения ключевых метрик

Шаблон для аудита базы знаний и контента для обучения ИИ

Детализированный аудит-лист для оценки качества и полноты существующей базы знаний, а также выявления пробелов для эффективного обучения ИИ-ассистента.

Общая информация о базе знаний (источники, актуальность)

- ☐ Определены ли все источники данных для обучения ИИ?
Это могут быть внутренние базы данных, логи чатов, FAQ, инструкции, регламенты, записи звонков, открытые репозитории (Kaggle, UCI Machine Learning Repository) или API сторонних сервисов (Parsingmaster.com).
- ☐ Сохранены ли «сырые» данные для аудита и воспроизводимости?
Важно иметь исходные данные «как есть» перед любой обработкой (Parsingmaster.com).
- ☐ Исключены ли персональные данные (ПДн) и чувствительные поля, если они не требуются для задачи?
Если задача решается без контактов/профилей/адресов — лучше не тащить их в датасет вовсе (Parsingmaster.com).
- ☐ Есть ли четкое понимание цели использования ИИ и минимального набора полей?
Избегайте ошибки «собрать всё, вдруг пригодится» (Parsingmaster.com).
- ☐ Определено ли происхождение данных (provenance)?
Источник (домен/система/выгрузка), условия использования, дата сбора. Это нужно для расследований, репликации эксперимента и юридической чистоты (Parsingmaster.com).
- ☐ Назначен ли ответственный за регулярное обновление материалов?
База знаний не статична: появляются новые продукты, меняются условия, клиенты задают новые вопросы.

Структура и навигация

- ☐ Насколько логично иерархически структурирована база знаний?
Наличие категорий, подкатегорий, тегов.
- ☐ Легко ли найти ответы на 80% типовых вопросов?
Оцените скорость поиска информации для оператора или пользователя.
- ☐ Есть ли удобная система поиска по базе знаний?
Эффективность поиска по ключевым словам, фразам.
- ☐ Используются ли перекрестные ссылки между связанными статьями?
Помогает связать информацию и улучшить понимание.
- ☐ Применяется ли единая терминология и глоссарий?
Устраняет путаницу и обеспечивает однозначность.

Качество контента (полнота, точность, язык)

- ❑ Все ли ответы содержат однозначное решение или четкий путь к нему?
Избегайте двусмысленности.
- ❑ Используется ли единый стиль и тон?
Соответствие брендбуку и корпоративной культуре.
- ❑ Насколько полно раскрыта каждая тема?
Достаточно ли информации для полного понимания вопроса без дополнительных источников.
- ❑ Содержит ли контент примеры, иллюстрации, скриншоты, где это уместно?
Визуализация улучшает понимание.
- ❑ Проверена ли информация на грамматические, орфографические и пунктуационные ошибки?
Качество текста напрямую влияет на доверие и понимание.
- ❑ Понятен ли язык изложения для целевой аудитории?
Избегайте излишнего жаргона или слишком сложной терминологии.
- ❑ Соответствует ли контент требованиям к качеству данных для ИИ?
Модель почти всегда «наследует» качество данных. Если датасет шумный, смещенный, с утечками — результат будет хрупким (Parsingmaster.com).

Актуальность данных

- ❑ Есть ли устаревшая информация?
Например, ссылки на неактуальные продукты, услуги, акции, контактные данные.
- ❑ Проводится ли регулярный аудит актуальности контента?
Например, ежеквартально или при каждом изменении продукта/услуги.
- ❑ Есть ли механизм для оперативного обновления информации при изменениях?
Особенно критично для динамично меняющихся продуктов или услуг.
- ❑ Все ли ссылки и внешние источники актуальны и доступны?
Проверка битых ссылок.

Полнота покрытия запросов

- ❑ Покрывает ли база знаний 80% наиболее частых запросов клиентов?
Анализ логов обращений, чатов, звонков для выявления типовых вопросов.
- ❑ Есть ли пробелы в ответах на важные, но менее частые вопросы?
ИИ-ассистент должен уметь отвечать на широкий спектр запросов.
- ❑ Учитываются ли различные формулировки одного и того же вопроса?
Например, 'как вернуть товар', 'возврат покупки', 'обменять товар'.
- ❑ Есть ли контент для поддержки клиентов на разных этапах жизненного цикла продукта/услуги?
От онбординга до решения проблем и продления.
- ❑ Проанализированы ли логи операторов за последние 6 месяцев для выявления повторяющихся паттернов?
Для обучения модели требуется минимум 1000 реальных диалогов.

Выявление дубликатов и противоречий

- ❑ Есть ли дублирующаяся информация в разных разделах или статьях?
Может привести к путанице и неэффективному использованию ресурсов.
- ❑ Содержатся ли противоречивые ответы или инструкции?

Критично для ИИ, так как он не сможет выбрать правильный ответ.

☐ Проведена ли дедупликация данных?

Удаление копий и «почти копий» (Parsingmaster.com). Автоматизируйте удаление дубликатов через `df.drop_duplicates(subset=['email'])` (Future-hub.io).

☐ Проведена ли очистка данных от мусора и ошибок формата?

Большинство реальных данных поступают в «сыром» виде: с пропусками, дубликатами, опечатками, устаревшими или некорректными значениями (Vc.ru).

Сложные и нестандартные запросы

☐ Определены ли типы запросов, которые ИИ не должен обрабатывать самостоятельно?

Например, жалобы, конфликтные ситуации, запросы, требующие эмпатии.

☐ Есть ли четкие правила маршрутизации таких запросов на живого оператора?

ИИ должен уметь передать диалог, если не может помочь.

☐ Обеспечивается ли передача контекста диалога оператору при эскалации?

Оператор должен видеть всю историю общения с ИИ.

☐ Созданы ли инструкции для операторов по работе со сложными запросами, которые могут быть частично обработаны ИИ?

ИИ-ассистент может предоставлять подсказки оператору.

Потенциал для автоматизации

☐ Выявлены ли процессы, где ИИ может значительно сократить время ответа или обработки?

Например, ответы на типовые вопросы, запись на услугу, подтверждение заказа, выдача справок.

☐ Есть ли данные для обучения ИИ для персонализации коммуникаций?

История покупок, частота обращений, поведение на сайте.

☐ Возможно ли использование ИИ для анализа настроений и эмоций клиентов?

Выявление негативных отзывов, адаптация тона коммуникации.

☐ Может ли ИИ помочь в предиктивной аналитике (прогнозирование нагрузки, оттока)?

Анализ исторической сезонности, рекламных кампаний, изменений в продукте.

☐ Есть ли возможность интеграции ИИ с существующими CRM-системами (Битрикс24, AmoCRM) или другими платформами?

Многие CRM уже включают ИИ-модули.

Критерии качества данных для обучения ИИ

Объем данных

Для начала с минимального объема, затем итеративный подход. Если метрики выходят на плато, а разрыв между тренировочными и тестовыми метриками невелик, данных достаточно (Vc.ru). Для обучения модели требуется минимум 1000 реальных диалогов.

Полнота

Отсутствие пропущенных значений. Если пропуски <5%, удалите строки; >20% — удалите столбец или генерируйте синтетические данные (Future-hub.io).

Точность

Отсутствие ошибок, опечаток, некорректных значений. Текстовые данные особенно чувствительны к сленгу, неверной кодировке (Vc.ru).

Репрезентативность	Данные должны отражать реальные сценарии использования и распределение запросов (Vc.ru).
Актуальность	Свежие данные, отражающие текущее состояние продуктов и услуг.
Согласованность	Единые форматы, терминология, отсутствие противоречий.
Отсутствие смещений (Bias)	Данные не должны содержать скрытых предубеждений, которые могут привести к несправедливым или некорректным ответам ИИ (Vc.ru).
Разделение данных	Разделите данные на три группы: 60% – обучение, 20% – валидация, 20% – тест. Для временных рядов сохраняйте хронологический порядок (Future-hub.io).

Оценка готовности базы знаний к обучению ИИ



Не готова, требуется значительная доработка | Требуется доработка | Начальная готовность | Высокая готовность | Максимальная, высокое качество

Отметьте, где находится ваша текущая база знаний по шкале готовности к обучению ИИ.

Больше интересного в нашем блоге



развивайте бизнес-мышление с нашими исследованиями

Получить пользу

