## **Отчет**

### 

GitHub - <https://github.com/Im-Vares/Document_Comparison_with_LLMs>

### **Постановка задачи**

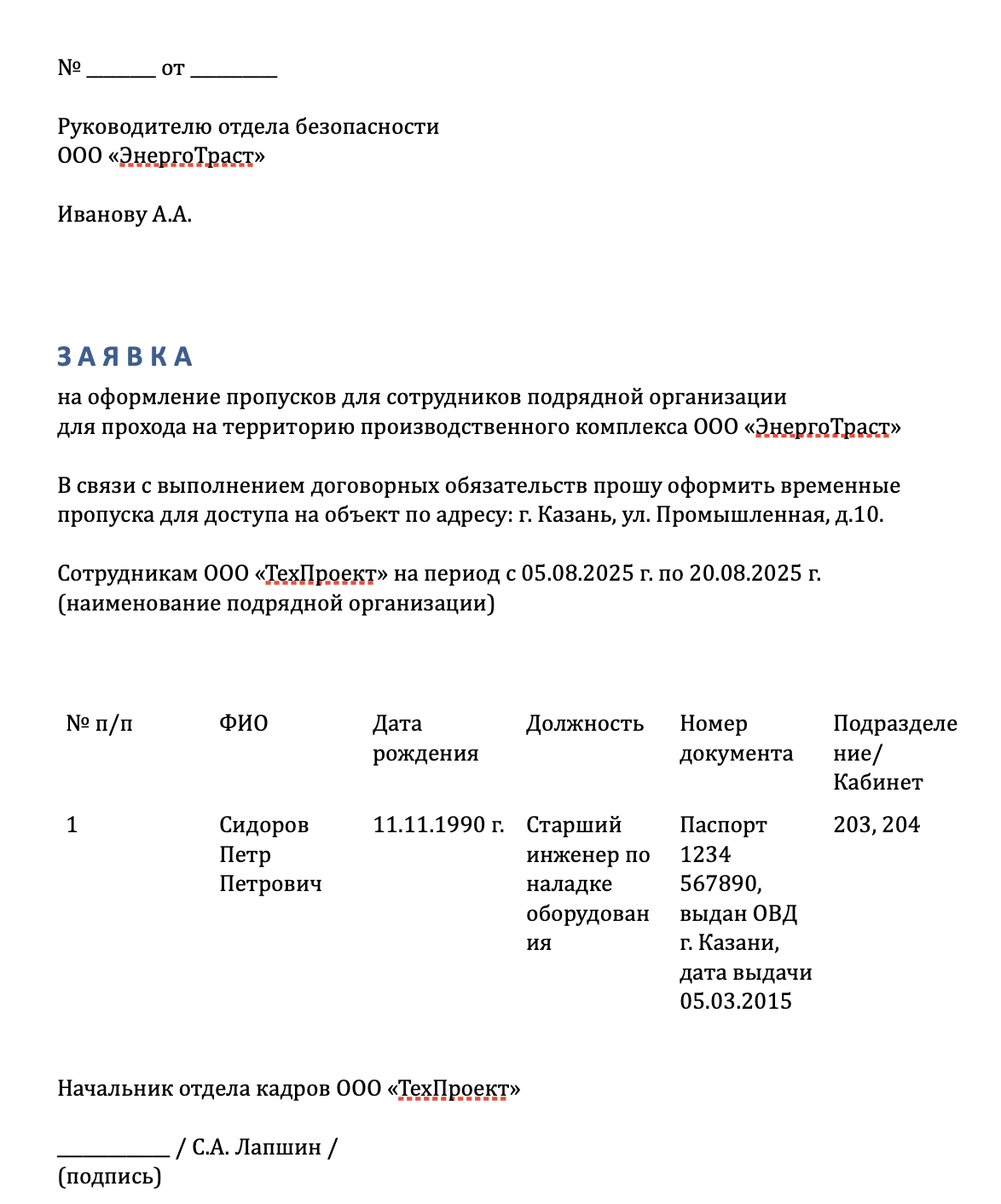
Целью исследования было определить оптимальный способ автоматизированного сравнения документов с использованием больших языковых моделей. В рамках работы необходимо было провести практическое тестирование трех основных методов — прямого сравнения текста, вопросно‑ответного подхода и двухшагового сравнения.

Задачи исследования включали проверку того, насколько каждый метод способен выявлять отсутствующие, измененные и лишние элементы документа, оценку удобства реализации и стабильности получаемых результатов, а также анализ сильных и слабых сторон каждого подхода. Дополнительно были изучены, но не тестировались в полном объеме, другие стратегии работы с текстами, такие как semantic diff, chunking, использование эмбеддингов и RAG, чтобы понять их потенциал для последующего применения. Итогом исследования должно было стать формирование выводов и рекомендаций относительно выбора метода, наиболее подходящего для решения задачи сравнения документов.

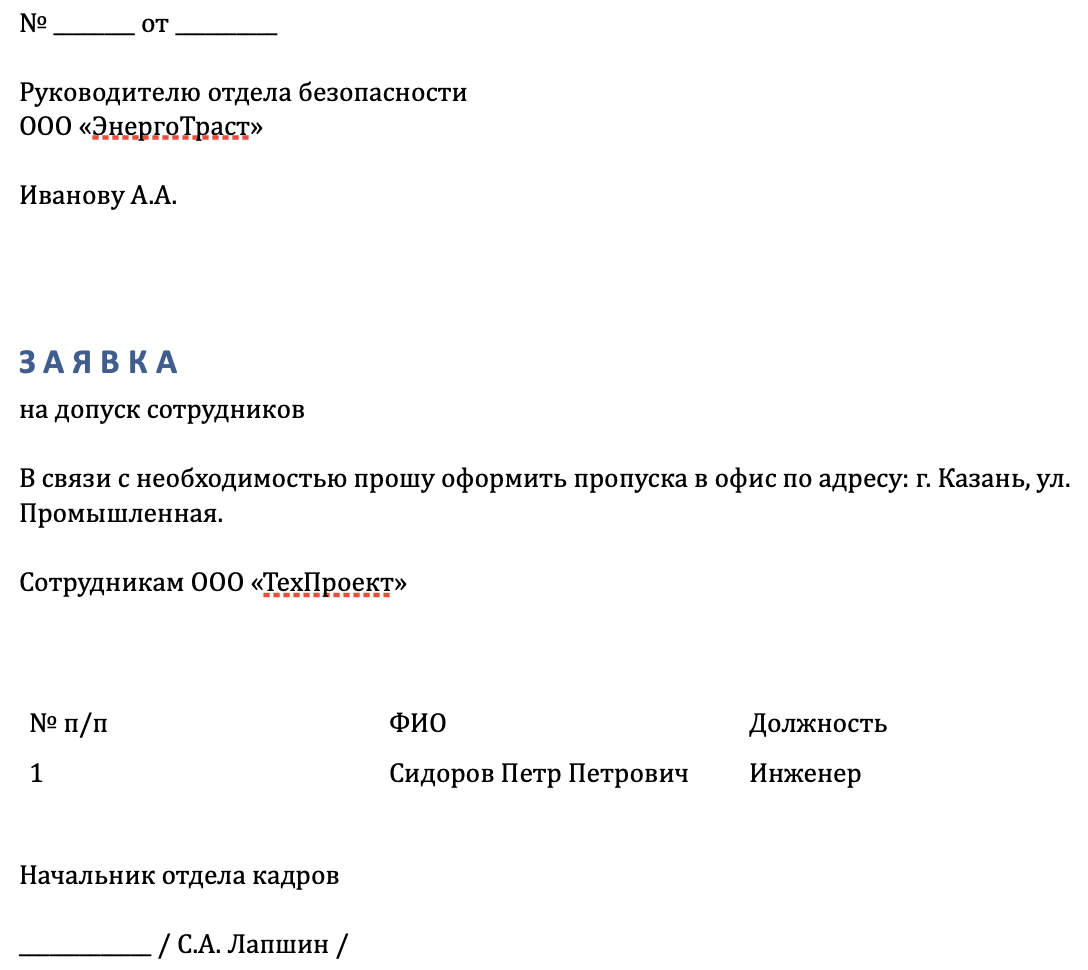
### **Описание подходов**

В ходе исследования было проведено тестирование трех методов сравнения документов с использованием LLM. Каждый метод решает задачу по‑своему и имеет как преимущества, так и ограничения.

Пример шаблона ТЗ:



Пример пользовательского ТЗ:



Список LLM:

* **LLaMA‑3‑8B** — meta-llama/Meta-Llama-3-8B-Instruct
* **Mistral‑7B** — mistralai/Mistral-7B-Instruct-v0.2
* **DeepSeek-R1** — deepseek-ai/DeepSeek-R1
* **Qwen3** — Qwen/Qwen3-Coder-480B-A35B-Instruct

**Первый метод** — прямое сравнение текста. Суть подхода заключалась в том, что шаблон и проверяемый документ полностью передавались модели с инструкцией выявить различия. Модель определяла, какие элементы отсутствуют, какие изменены и что было добавлено. Этот метод оказался самым простым в реализации: требуется только один запрос к LLM. Однако выявились недостатки. При большом объеме текста модель склонна терять структуру, смешивать детали и не всегда последовательно выделять все несоответствия.

**Промт:** Сравни два документа: эталонный и проверяемый. Выведи СТРОГО 3 раздела: ОТСУТСТВУЕТ: - что есть в эталоне, но нет во втором документе ИЗМЕНЕНО: - что было → что стало ДОБАВЛЕНО: - что есть во втором документе, но не было в эталоне

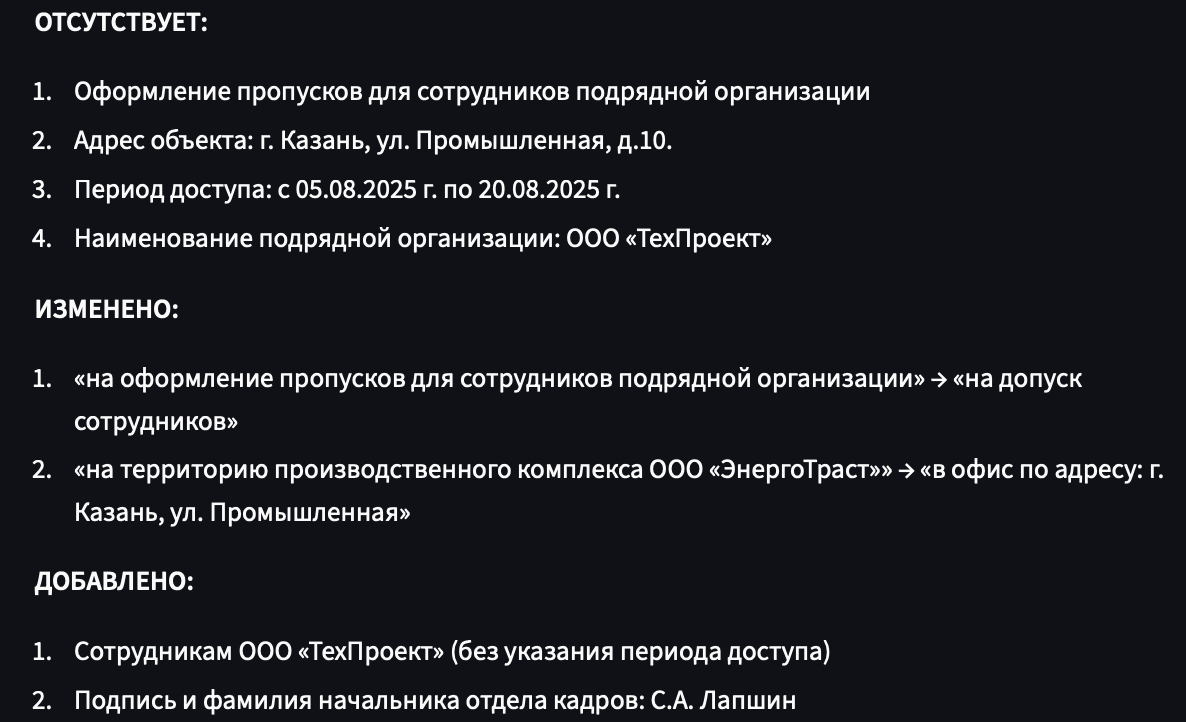


Рисунок 1 - Ответ Llama-3-8B

**Что сделано правильно:**

– Корректно указано отсутствие периода доступа «с 05.08.2025 г. по 20.08.2025 г.».

– Верно зафиксированы изменения формулировок: «на оформление пропусков…» → «на допуск сотрудников» и «на территорию производственного комплекса…» → «в офис по адресу…».

– Точно отмечено отсутствие полного адреса с номером дома.

**Что сделано плохо:**

– Ошибочно указано, что отсутствует наименование подрядной организации «ООО „ТехПроект“» — оно присутствует в прототипе.

– В категорию «Добавлено» попали элементы, которые уже есть в шаблоне: фраза «Сотрудникам ООО „ТехПроект“» не является новой, а подпись начальника отдела кадров не добавлена, она также есть в шаблоне.

– Не упомянуты другие отсутствующие детали (например, удалённые поля «Дата рождения», «Номер документа», «Подразделение/Кабинет»).

**Итоговая оценка:** ответ частично корректен — главные смысловые изменения отмечены, но присутствуют явные ошибки в классификации и неполный охват всех отличий.

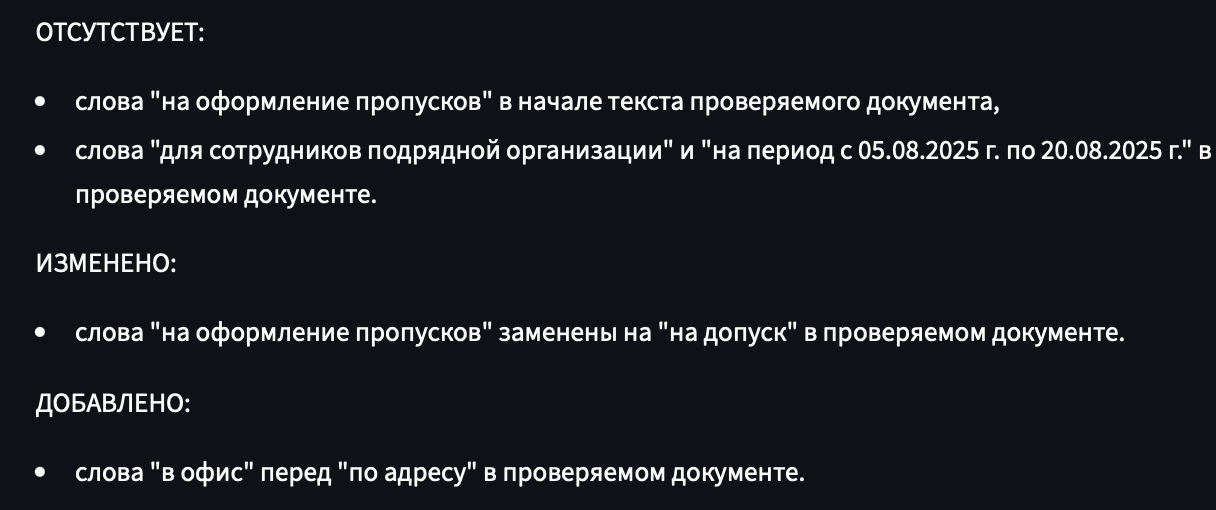
****

Рисунок 2 - Ответ Mistral‑7B

**Что сделано правильно:**

– Корректно указано отсутствие формулировки «на оформление пропусков».

– Верно зафиксировано отсутствие периода «с 05.08.2025 г. по 20.08.2025 г.».

– Правильно отмечено изменение «на оформление пропусков» → «на допуск».

– Верно указано добавление слов «в офис».

**Что сделано плохо:**

– Не упомянуто отсутствие фразы «на территорию производственного комплекса ООО „ЭнергоТраст“».

– Пропущено, что адрес неполный — в прототипе отсутствует номер дома «д.10».

– Не отмечено, что убрана деталь «для прохода на территорию», меняющая смысл.

– Слишком поверхностный разбор: многие элементы (например, подписи, структура таблицы, отсутствие граф «Дата рождения», «Номер документа») не были проанализированы.

**Итоговая оценка:** ответ частично отражает изменения, но анализ неглубокий и неполный. Модель указала лишь очевидные отличия и не рассмотрела другие существенные несоответствия между документами.

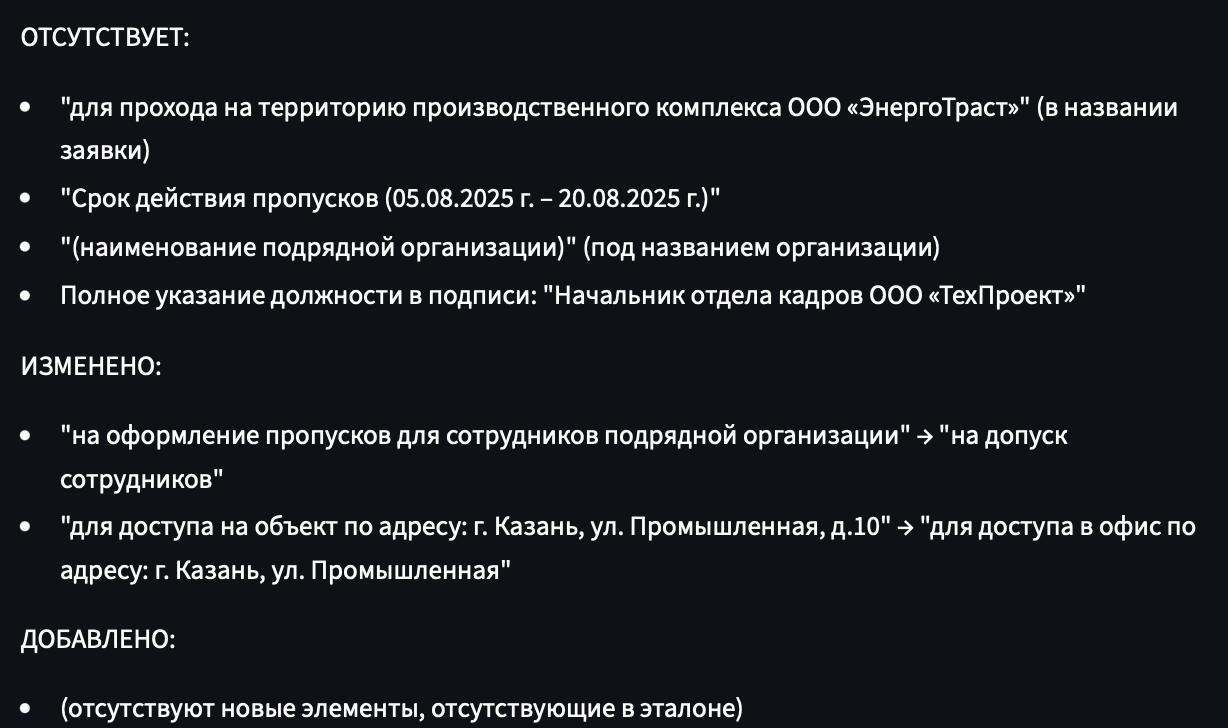


Рисунок 3 - Ответ DeepSeek-R1

**Что сделано правильно:**

– Верно выделены ключевые отсутствующие элементы: срок действия пропусков, полное название организации в подписи, указание “для прохода на территорию производственного комплекса ООО «ЭнергоТраст»”.

– Корректно зафиксированы смысловые изменения формулировок (“на оформление пропусков…” → “на допуск сотрудников”, “для доступа на объект…” → “для доступа в офис…”).

– Четко отмечено, что в проверяемом документе нет новых (добавленных) элементов.

**Что сделано плохо:**

– Указание “(наименование подрядной организации)” как отсутствующего спорно: в проверяемом документе “Сотрудникам ООО «ТехПроект»” присутствует, но нет отдельного пояснения “(наименование подрядной организации)” под этой строкой — модель трактует это как отсутствие, что можно считать чрезмерно буквальной придиркой.

– Не отражены другие явные пропуски, например, отсутствие полей таблицы “Дата рождения”, “Номер документа”, “Подразделение/Кабинет”.

– Нет анализа структуры документа (таблица урезана, несколько граф исчезли).

**Итоговая оценка:** ответ в целом аккуратен и правильно фиксирует главные смысловые различия, но неполный: не учтены структурные потери и сокращения данных в таблице.

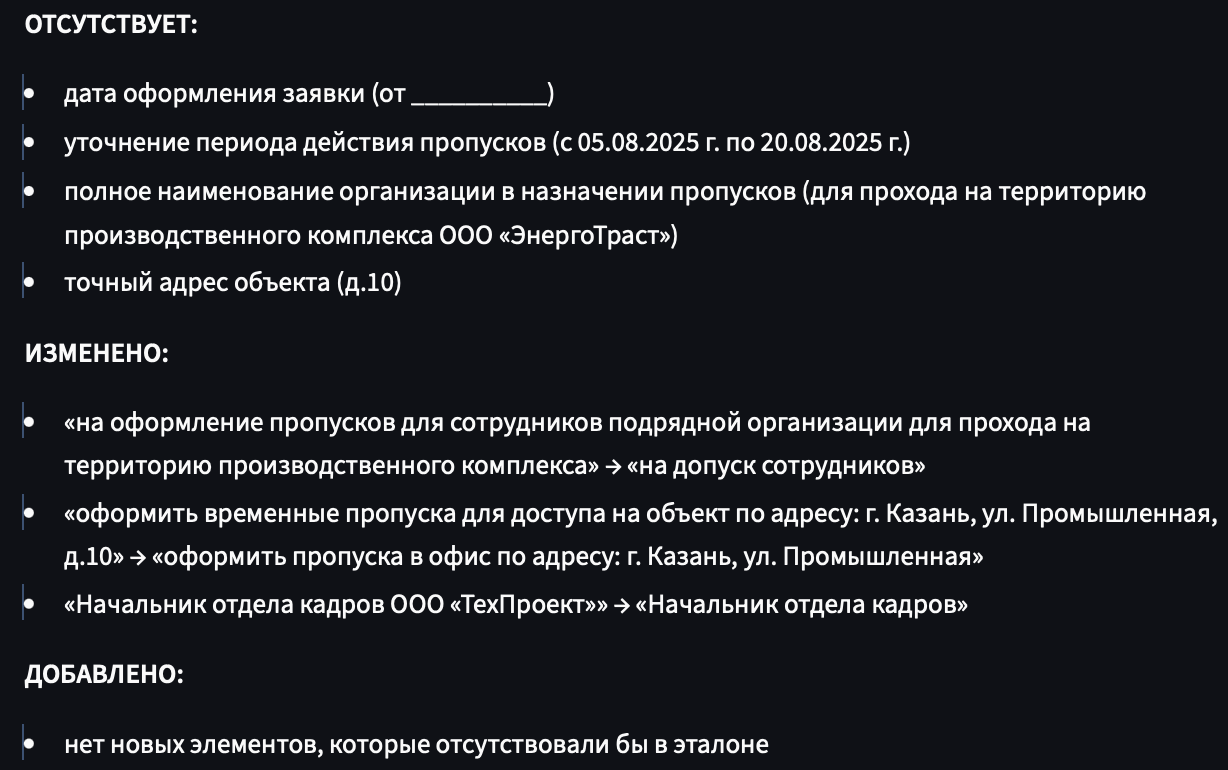


Рисунок 4 - Ответ Qwen3

**Что сделано правильно:**

– Верно определены ключевые отсутствующие элементы: дата оформления заявки, период действия пропусков, номер дома в адресе, упоминание полного наименования организации в назначении пропусков.

– Корректно зафиксированы основные изменения формулировок: название заявки, адрес назначения, сокращение подписи.

– Четко указано, что новых элементов нет, что соответствует действительности.

**Что сделано плохо:**

– Не упомянуто отсутствие полей таблицы (дата рождения, номер документа, подразделение/кабинет), что является существенным.

– Формулировка «полное наименование организации в назначении пропусков» звучит не совсем точно — в шаблоне речь о «территории производственного комплекса ООО «ЭнергоТраст»», а в отчете это обозначено несколько расплывчато.

– Недостаточно внимания к тому, что в прототипе сокращена сама структура документа: таблица урезана, несколько граф исчезли.

**Итоговая оценка:** ответ в целом точный и отражает основные смысловые изменения, но неполный: модель не учла структурные потери и часть деталей таблицы.

Для первого метода лучше всего использовать **Qwen3** – она наиболее точно держит структуру ответа, корректно выделяет категории и не теряет смысловые детали.

**Второй метод** — вопросно‑ответный подход. Здесь по каждому пункту шаблона формировались вопросы, которые последовательно задавались проверяемому документу. Такой метод позволял контролировать проверку пошагово: если элемент найден — фиксировался результат, если нет — отмечалось отсутствие. Преимущество этого подхода в том, что он дает более детальную и прозрачную картину по каждому элементу документа. Минус в том, что метод требует больше шагов и, соответственно, больше времени и ресурсов.

**Промт на генерацию вопросов:** “Составь 15+ проверочные вопросы, чтобы проверить соответствие прототипа эталонному документу. Вопросы должны охватывать: - наличие обязательных реквизитов (адрес, даты, сроки действия, компания, должности); - соответствие формулировок (например, «территория производственного комплекса» vs «офис»); - наличие всех таблиц и обязательных колонок (ФИО, должность, дата рождения, паспортные данные); - правильность оформления (подпись, формат заголовков, структура). Формат: нумерованный список. Эталонный текст:”

**Промт для проверки по вопросам:** “Отвечай на каждый вопрос на основе текста прототипа. Для каждого вопроса: - пиши «есть» или «нет» в начале ответа, - дай короткое объяснение (1-2 предложения), что именно есть или чего не хватает. Прототип: {example\_text} Вопросы: {questions\_text} Формат ответа: нумерованный список с пометками и пояснениями.”

**Промт для генерации итогового ответа:** “На основе ответов по вопросам составь единый отчет с инструкцией по исправлению прототипа. Раздели отчет на три блока: ЧТО ДОБАВИТЬ, ЧТО ИЗМЕНИТЬ, ЧТО УБРАТЬ. Для каждого пункта: - кратко объясни, почему это нужно; - предложи пример текста, как это должно выглядеть. Ответы по вопросам: {answers\_text} Формат: списки под каждым блоком.”

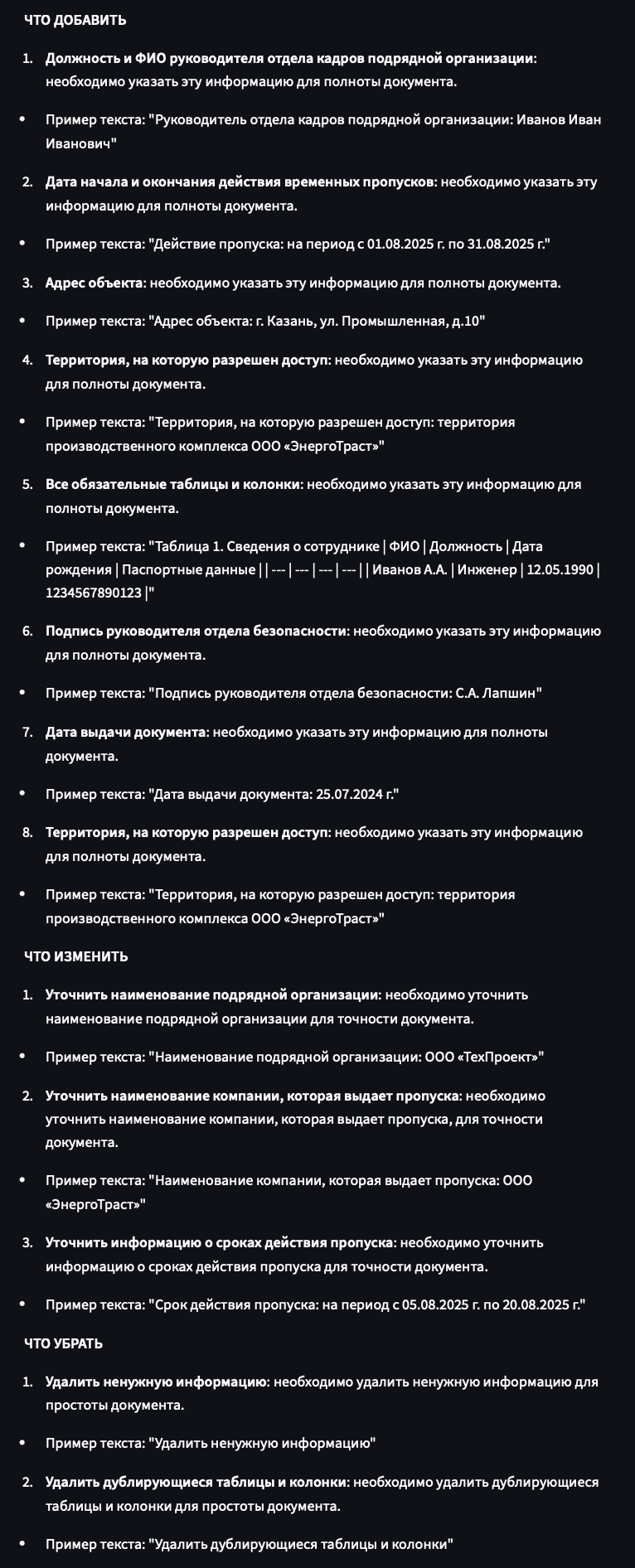


Рисунок 5 - LLaMa-3-8b

**Что сделано правильно:**

– Чёткая структура «ЧТО ДОБАВИТЬ / ЧТО ИЗМЕНИТЬ / ЧТО УБРАТЬ» с пояснениями и примерами для каждого пункта.

– Выявлены важные недостающие элементы (дата выдачи, сроки действия пропуска, адрес объекта, таблицы с данными сотрудников).

– Указаны корректные примеры формулировок, которые можно использовать прямо в документе.

**Что сделано плохо:**

– Есть дублирование: пункт «Территория, на которую разрешен доступ» повторён дважды.

– В разделе «ЧТО УБРАТЬ» рекомендации слишком общие и не содержат конкретных указаний, что именно нужно удалить (например, какие «ненужные разделы» или какие «дублирующиеся таблицы»).

– Формулировки из раздела «ЧТО УБРАТЬ» выглядят поверхностно и слишком универсально, не привязаны к конкретному содержимому прототипа.

**Итоговая оценка:**

Отчёт охватывает ключевые моменты и даёт понятные рекомендации, но требует доработки: убрать дублирующиеся пункты, уточнить, что именно подразумевается под «ненужной информацией» и «дублирующимися таблицами», чтобы документ был не только структурированным, но и конкретным.

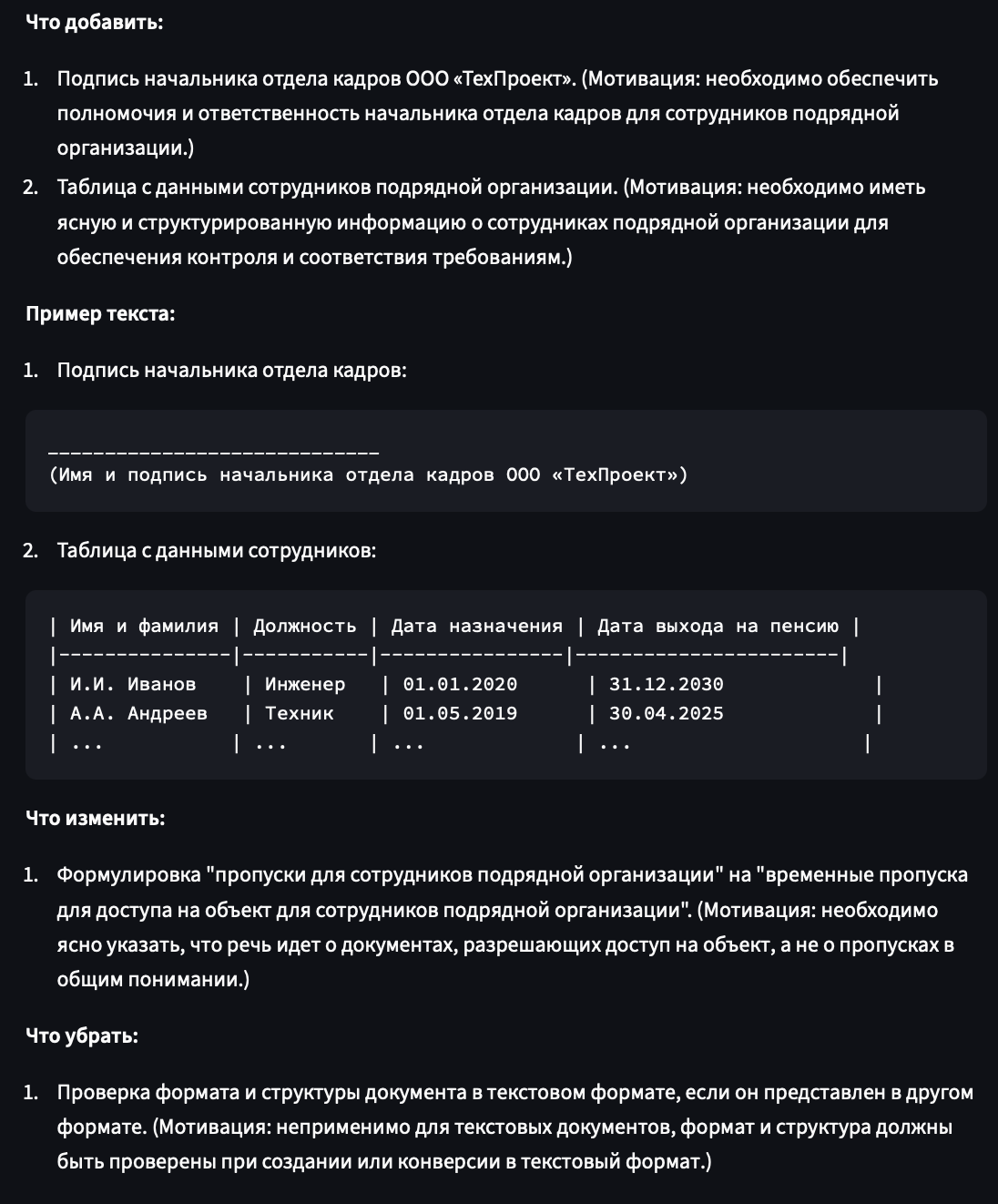


Рисунок 6 - Mistral‑7B

**Что сделано правильно:**

– Чёткая структура: разделы «Что добавить», «Что изменить», «Что убрать» оформлены последовательно.

– Указаны конкретные элементы, которые действительно нужны для полноты документа (подпись начальника отдела кадров, таблица с данными сотрудников).

– Есть пояснения, зачем требуется каждое дополнение или изменение, и приведены примеры текста, которые можно сразу вставить в документ.

**Что сделано плохо:**

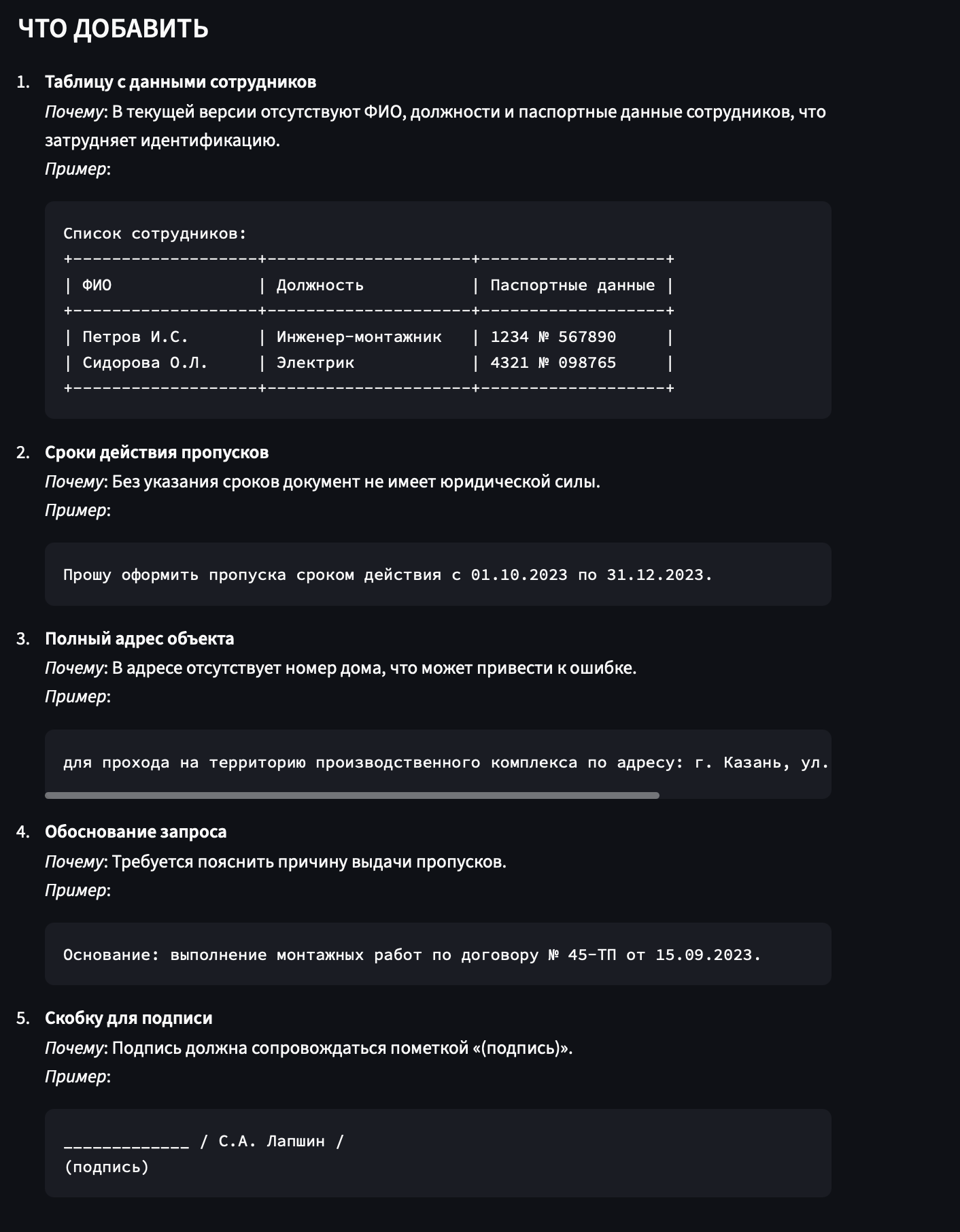
– Формулировка «Проверка формата и структуры документа…» в разделе «Что убрать» слишком общая и оторвана от контекста — непонятно, что именно предлагается удалить из прототипа.

– В разделе «Что изменить» представлен только один пункт — это делает отчёт несбалансированным: лучше было бы перечислить все найденные формулировки, требующие уточнения.

– Не учтены другие важные отсутствующие элементы (например, паспортные данные, сроки действия пропуска), хотя они логично вытекают из предыдущих версий отчёта.

**Итоговая оценка:**

Отчёт полезный и содержит конкретные, применимые рекомендации, но слишком узкий по охвату: требуется добавить больше уточнений по изменяемым элементам и конкретизировать, что именно подразумевается под удалением «проверки формата и структуры документа».



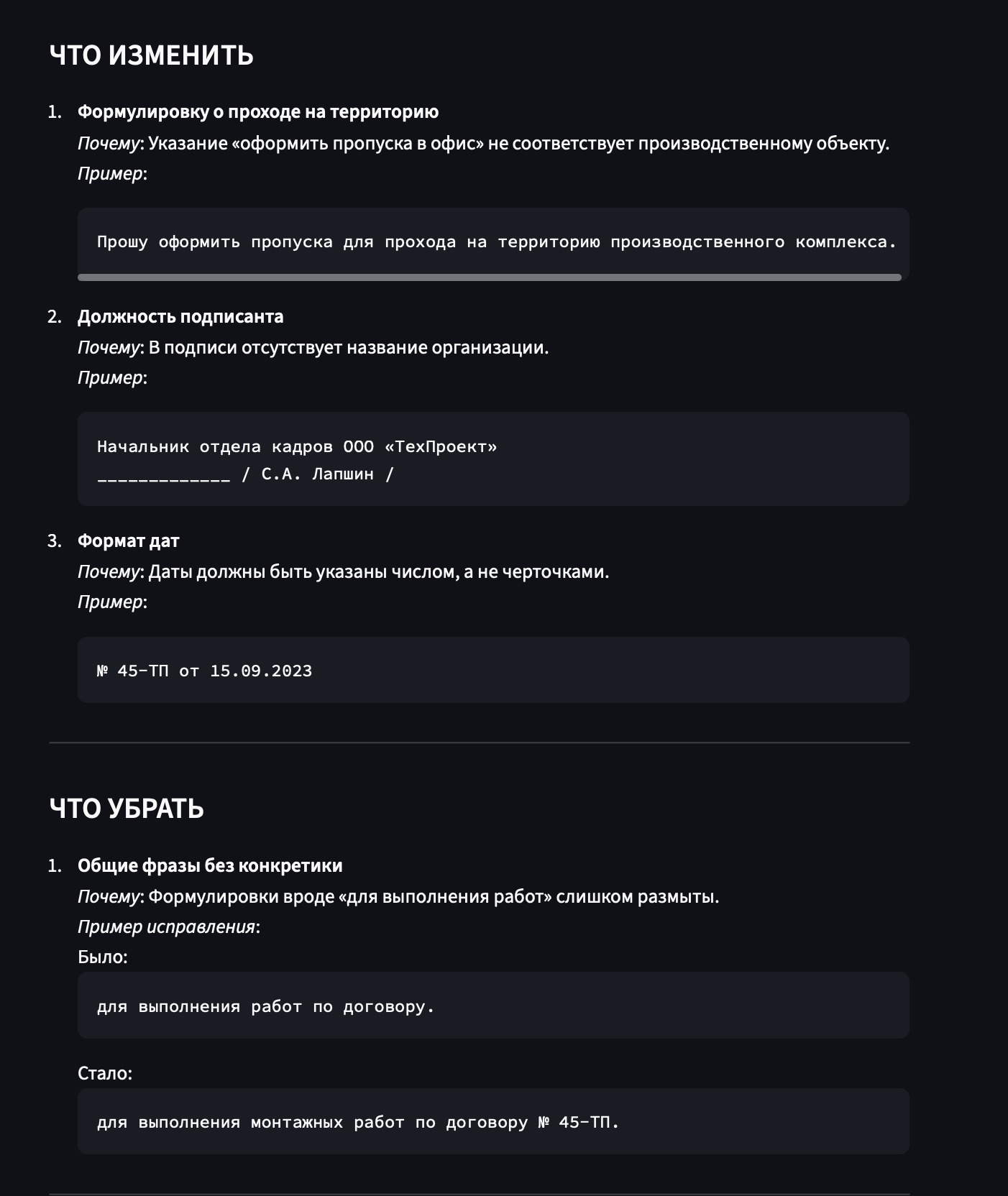


Рисунок 6-7 - DeepSeek-R1

**Что сделано правильно:**

– Чёткая структура на три блока: «Что изменить», «Что убрать», «Что добавить»; для каждого пункта даны пояснения «почему» и примеры правильного текста.

– В разделе «Что добавить» охвачены важные элементы: таблица сотрудников, сроки действия пропусков, полный адрес, обоснование запроса, оформление подписи.

– В разделе «Что изменить» корректно выделены формулировки, требующие доработки (проход на территорию, должность подписанта, формат дат).

– Показана разница между текущей формулировкой и исправленным вариантом (например, общий текст «для выполнения работ» заменён на конкретизированный).

**Что сделано плохо:**

– Есть пересечение и дублирование пунктов (например, в блоке «Что добавить» некоторые элементы можно объединить — “полный адрес объекта” и “формулировка прохода на территорию” логично свести в один пункт).

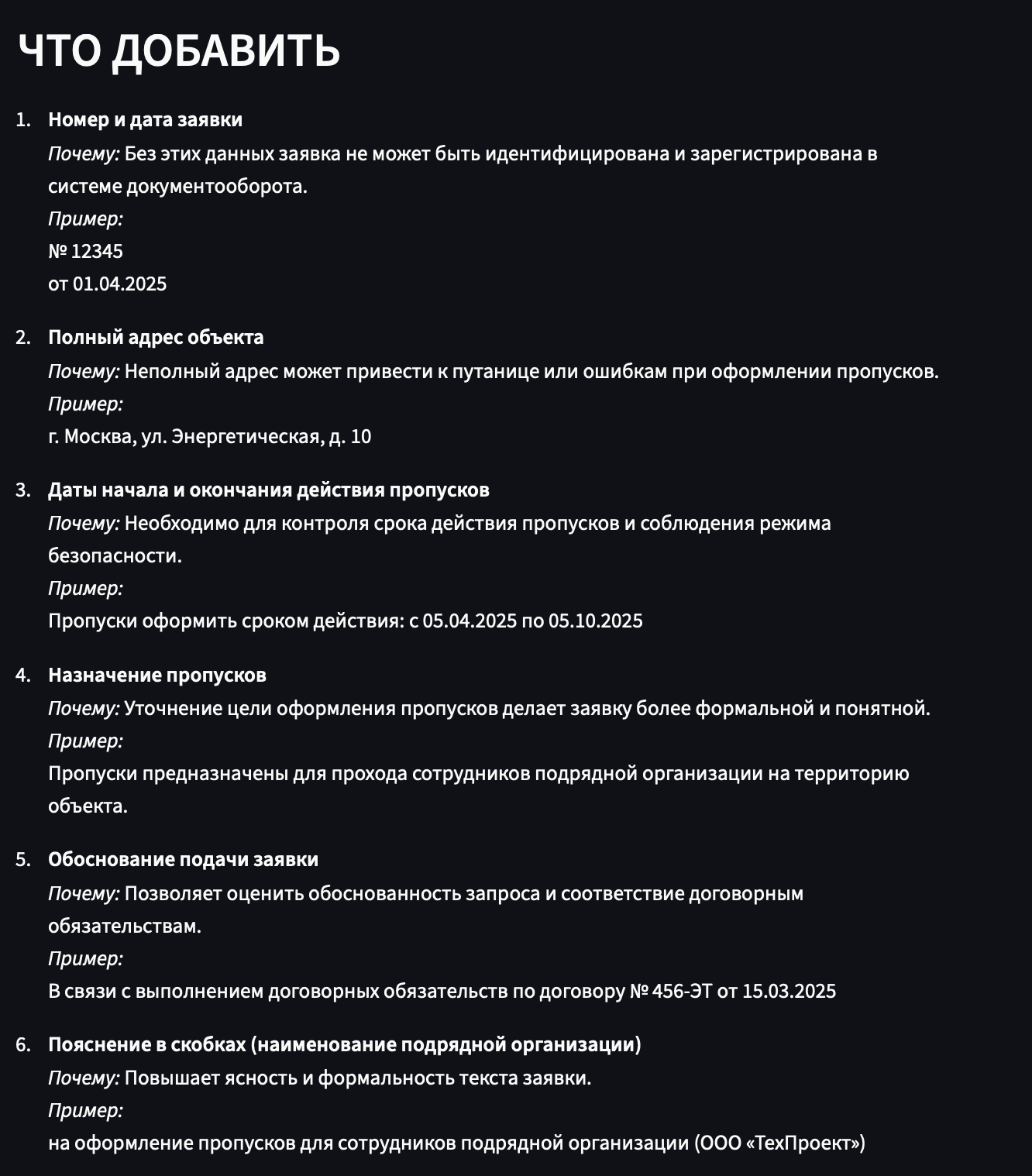
– В разделе «Что убрать» слишком общая рекомендация “удалить ненужную информацию” без привязки к конкретным частям прототипа; требует конкретизации.

– Структура таблицы сотрудников показана, но в примере указаны только две строки — отчёт мог бы пояснить, что таблица должна быть расширяемой и содержать все обязательные колонки.

– Не отражены изменения, которые связаны с отсутствием дополнительных полей (например, «подразделение/кабинет»), хотя они есть в эталонном документе.

**Итоговая оценка:**

Рекомендации системные и чётко оформленные, но их нужно доработать: убрать дубли, конкретизировать, что именно “ненужное” удаляется, и расширить перечень исправлений для полноты охвата всех отличий между эталоном и прототипом.



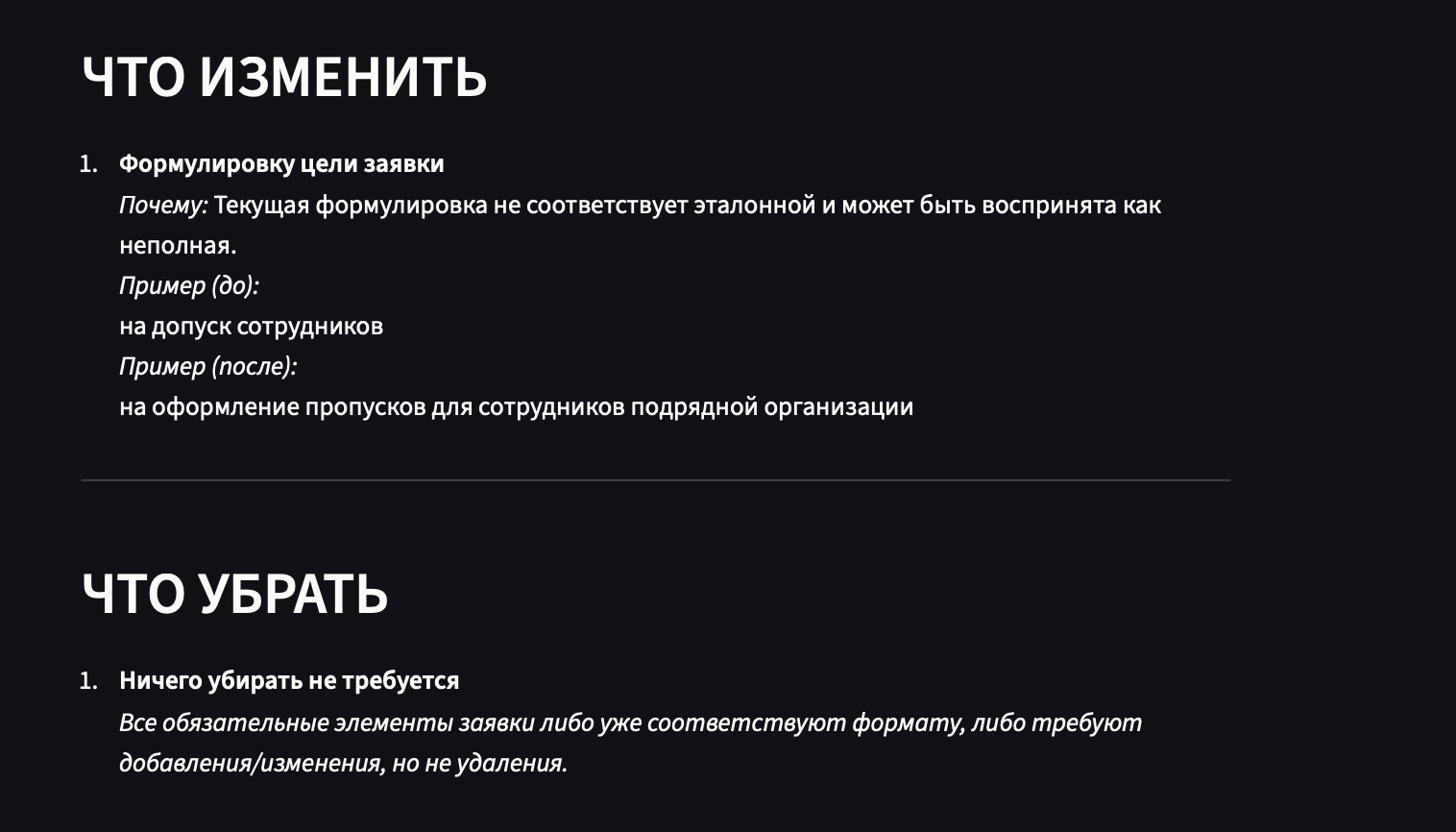


Рисунок 8-9 - Qwen3

**Что сделано правильно:**

– Рекомендации чётко разделены на три раздела, логика структуры соблюдена.

– В разделе «Что добавить» перечислены действительно важные недостающие элементы (номер и дата заявки, полный адрес объекта, даты действия пропусков, назначение и обоснование заявки).

– Для каждого пункта есть пояснение «почему» и пример текста, что облегчает внедрение рекомендаций.

– Раздел «Что убрать» лаконичен и сразу снимает вопрос о ненужных элементах.

**Что сделано плохо:**

– В разделе «Что изменить» указан только один пункт (формулировка цели заявки) — список выглядит неполным, хотя можно было бы добавить уточнение формата подписи или дат.

– Некоторые формулировки в «Что добавить» дублируют друг друга по сути: «назначение пропусков» и «обоснование подачи заявки» можно объединить в один блок для большей чёткости.

– Не затронут вопрос о корректности таблицы с данными сотрудников и отсутствующих полей (дата рождения, паспортные данные и т.п.), что важно для соответствия эталону.

**Итоговая оценка:**

Отчёт структурирован и в целом полезен — он закрывает ключевые пробелы и не предлагает лишнего. Однако он не полностью охватывает все расхождения и может быть дополнен деталями по структуре документа и формату отдельных элементов, чтобы рекомендации были исчерпывающими.

DeepSeek‑R1 выдал самый полный и системный отчёт: он охватил все ключевые элементы документа, поделил рекомендации на «изменить», «убрать» и «добавить» и дал конкретные формулировки с пояснениями. Другие модели либо писали слишком узко, либо уходили в общие фразы.

**Третий метод** — двухшаговое сравнение. Этот подход строился на двойной проверке. Сначала анализировался шаблон: что в нем есть и чего нет в проверяемом документе. Затем анализировался сам прототип: что в нем добавлено сверх шаблона.

**Промт:** “Ты – аудитор документов. Сравни шаблон (эталон) и проверяемый вариант. Сделай два шага анализа: 1. Template → Prototype - Определи, какие элементы шаблона полностью отсутствуют (пометь как 'Отсутствует'). - Определи, какие элементы присутствуют, но частично (пометь как 'Частично'). 2. Prototype → Template - Найди, что добавлено в проверяемый документ, но отсутствует в шаблоне (пометь как 'Лишнее'). Ответ должен содержать три раздела: ОТСУТСТВУЕТ: - ... ЧАСТИЧНО: - ... ЛИШНЕЕ: - ... --- ШАБЛОН: {template\_text} --- ПРОВЕРЯЕМЫЙ: {example\_text} --- Ответ:”

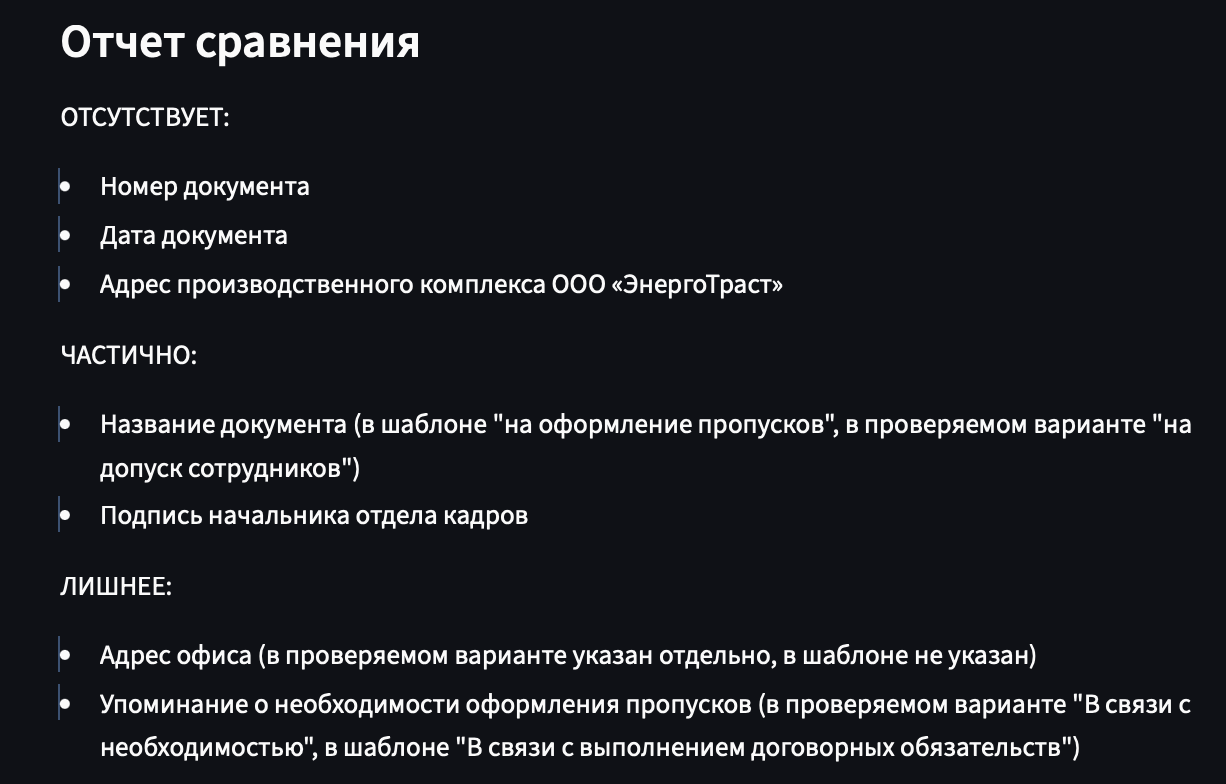


Рисунок 10 - LLaMa-3-8b

**Что сделано правильно:**

– Верно определены отсутствующие ключевые элементы: номер документа, дата документа, адрес производственного комплекса.

– Корректно отмечено, что название документа и подпись начальника отдела кадров даны частично (сокращения и упрощения по сравнению с шаблоном).

– Обнаружено, что в проверяемом документе есть элементы, которых нет в шаблоне (адрес офиса, фраза «В связи с необходимостью»).

**Что сделано плохо:**

– Формулировка «упоминание о необходимости оформления пропусков» как “лишнее” спорна: это не новый блок текста, а изменение формулировки («в связи с выполнением договорных обязательств» → «в связи с необходимостью»), что скорее относится к категории «ЧАСТИЧНО».

– Не отмечены другие пропуски из таблицы (дата рождения, паспортные данные, подразделение/кабинет).

– Категории неравномерно проработаны: в «Лишнее» отмечены общие фразы, но не конкретизировано, в каких местах текста они встречаются.

**Итоговая оценка:**

Ответ LLama по третьему подходу в целом корректно выделяет основные отсутствующие и частично заполненные элементы, но допускает ошибки в классификации (часть “лишнего” относится к изменённым формулировкам) и неполно охватывает детали структуры документа.

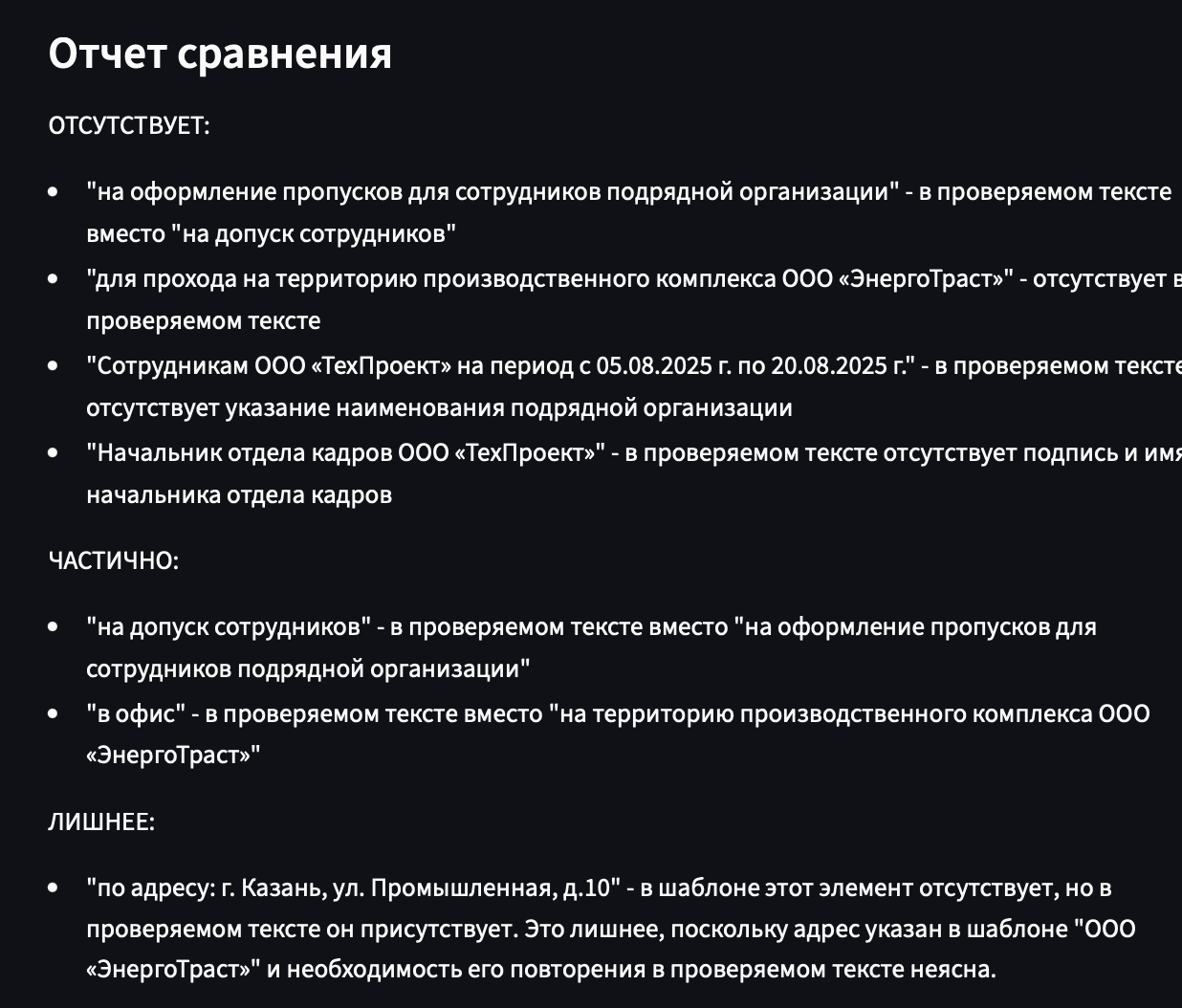


Рисунок 11 - Mistral-7b

**Что сделано правильно:**

– Правильно указано, что отсутствует ключевая формулировка «на оформление пропусков для сотрудников подрядной организации».

– Верно отмечено, что нет блока «для прохода на территорию производственного комплекса ООО „ЭнергоТраст“».

– Корректно зафиксировано, что подпись «Начальник отдела кадров ООО „ТехПроект“» отсутствует.

– Раздел «Частично» отражает реальные замены («на допуск сотрудников» и «в офис»).

**Что сделано плохо:**

– Фраза «Сотрудникам ООО „ТехПроект“ на период с 05.08.2025 г. по 20.08.2025 г.» описана некорректно: в проверяемом тексте указание «Сотрудникам ООО „ТехПроект“» есть, отсутствует именно период, а не всё наименование подрядной организации.

– Элемент «по адресу: г. Казань, ул. Промышленная, д.10» неверно классифицирован как «лишний». Адрес должен присутствовать, но в шаблоне он оформлен иначе; это скорее **изменение**, а не добавление «лишнего».

– Не отмечены отсутствующие данные из таблицы (дата рождения, должность, паспортные данные), что является серьёзной потерей информации.

**Итоговая оценка:**

Ответ Mistral фиксирует ключевые смысловые расхождения, но неверно классифицирует часть элементов (адрес помечен как «лишний», а не изменённый), а также упускает важные пропуски в структуре документа.

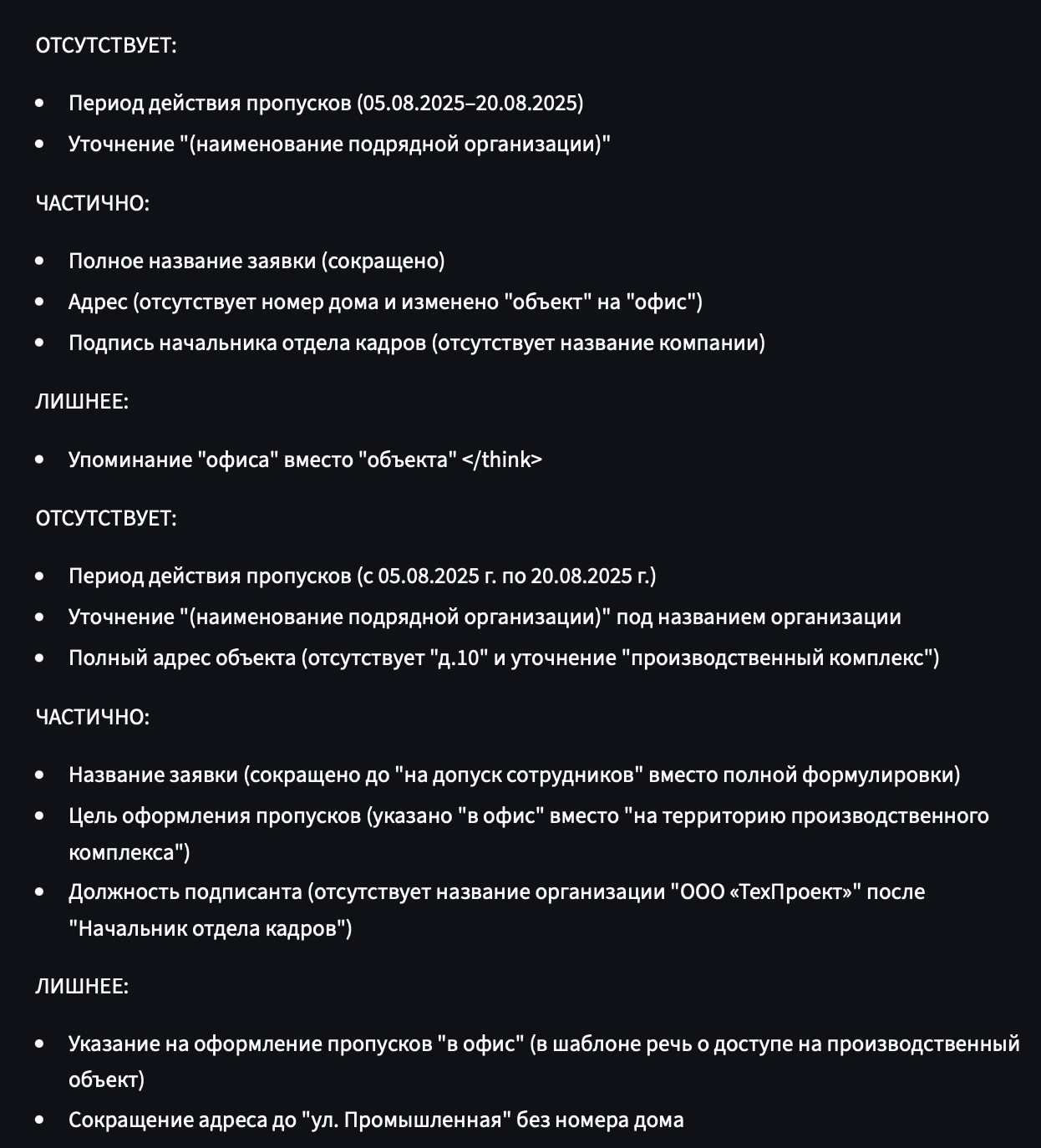


Рисунок 12 - DeepSeek-R1

**Что сделано правильно:**

– Правильно указаны отсутствующие данные: период действия пропусков, уточнение «(наименование подрядной организации)».

– Корректно зафиксированы частичные несоответствия: сокращение названия заявки, неполный адрес и подпись без упоминания организации.

– Логично выделено, что формулировка «в офис» не совпадает с шаблоном, и отмечено как отдельный пункт.

**Что сделано плохо:**

– В разделе «Отсутствует» пункт «Полный адрес объекта» описан так, будто он отсутствует полностью, хотя адрес есть, но изменён (корректнее классифицировать часть как «Частично»).

– Формулировка «Указание на оформление пропусков “в офис”» в «Лишнее» спорна: это скорее **изменение формулировки**, а не абсолютно новый элемент.

– Не упомянуты другие пропущенные детали таблицы (дата рождения, паспортные данные, подразделение/кабинет), которые есть в шаблоне, но отсутствуют в проверяемом тексте.

**Итоговая оценка:**

Ответ DeepSeek‑R1 хорошо фиксирует ключевые расхождения и корректно распределяет большинство элементов по категориям, но в классификации отдельных пунктов есть неточности, и охват структуры документа неполный.

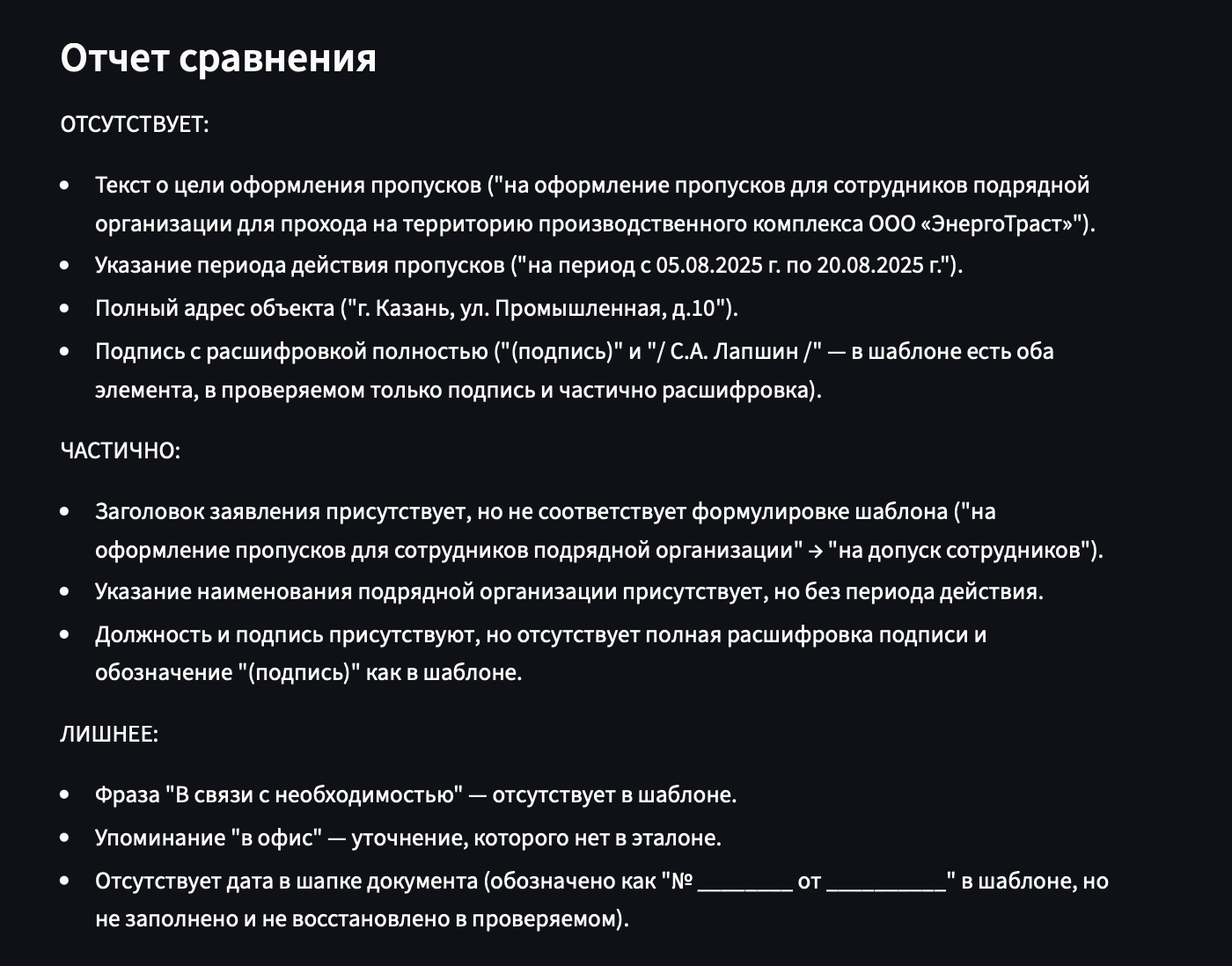


Рисунок 13 - Qwen3

**Что сделано правильно:**

– Чётко зафиксированы отсутствующие ключевые элементы: цель оформления пропусков в полном виде, период действия, полный адрес объекта, корректная подпись с расшифровкой.

– Раздел «Частично» грамотно выделяет сокращения и неполные формулировки (заголовок, упрощённая подпись, отсутствие периода действия).

– В «Лишнее» корректно упомянуты фразы, которых нет в шаблоне («В связи с необходимостью», уточнение «в офис»).

**Что сделано плохо:**

– Пункт «Отсутствует дата в шапке документа» ошибочно помещён в раздел «Лишнее»: это не лишний элемент, а наоборот — недостающий (дата не указана).

– Не упомянуты недостающие поля из таблицы (дата рождения, паспортные данные, подразделение/кабинет).

– Некоторые формулировки звучат неаккуратно: «Подпись с расшифровкой полностью» объединяет несколько проблем, но классифицирована как единый пункт — лучше разнести по отдельным аспектам (отсутствие пометки «(подпись)» и неполная расшифровка).

**Итоговая оценка:**

Ответ Qwen3 детально фиксирует смысловые расхождения и основные пропуски, но содержит ошибку в классификации даты и не охватывает все структурные недочёты (особенно таблицу). В целом это наиболее аккуратный отчёт из всех, но требует уточнений по структуре и распределению пунктов.

Qwen3 показал наилучший результат, потому что даёт самый полный и аккуратный отчёт: он фиксирует ключевые отсутствующие элементы, отмечает частичные несоответствия и выделяет лишние фразы без серьёзных пропусков.

### **Статьи**

**Semantic diff**

Semantic diff — это подход, при котором сравнение документов проводится не по буквам, а по смыслу текста. LLM определяет, совпадает ли значение фразы, даже если формулировка другая. Это позволяет игнорировать мелкие перестановки и синонимы, фиксируя только смысловые потери или искажения. Основное преимущество метода — точность в выделении «реальных» отличий, а главный недостаток — необходимость чёткой настройки критериев, иначе модель может трактовать одно и то же выражение по‑разному.

**Источник:** статья *“3 Approaches to Semantic Document Comparison”* (Medium, 2024) [ссылка](https://medium.com/@angela.shi/3-approaches-to-semantic-document-comparison-with-a-llm-2b6434dd2dd6)

**Chunking + Summarization**

Chunking + Summarization делит документ на небольшие части (чанки), делает резюме для каждого фрагмента и уже затем сравнивает эти резюме. Такой подход снижает нагрузку на модель и подходит для анализа длинных документов. Преимущество — возможность работать с большими текстами, недостаток — риск потери деталей при суммаризации.

**Источник:** блог Justin’s Blog *“Document Summarisation and Comparison with LLMs”* (2024) [ссылка](https://justinmatters.co.uk/wp/document-summarisation-and-comparison-with-llms/)

**Embeddings + LLM**

В этом подходе сначала создаются эмбеддинги (векторные представления) для каждого блока документа, и рассчитывается их сходство (cosine similarity). LLM анализирует только участки с низким сходством, объясняя, чем они отличаются. Это экономит токены и ускоряет процесс. Плюсы — масштабируемость и эффективность, минусы — необходимость инфраструктуры для эмбеддингов и векторной БД.

**Источник:** статья Codesphere *“AI Document Comparison for Compliance Using Embeddings”* (2025) [ссылка](https://huggingface.co/blog/g-ronimo/semscore)

**RAG**

RAG подразумевает, что документы индексируются в базе (например, FAISS), и модель при анализе запрашивает только релевантные фрагменты. Это делает возможным сравнение даже огромных архивов. Метод удобен, если нужно многократно сверять разные версии документов с шаблоном. Плюсы — работа с любыми объёмами, минусы — сложность настройки и риск смешивания данных, если retrieval настроен неправильно.

**Источник:** обсуждение на Reddit *“Document comparison RAG, the struggle is real”* (r/LocalLLaMA, 2024) [ссылка](https://www.reddit.com/r/LocalLLaMA/comments/1cn659i/document_comparison_rag_the_struggle_is_real/)

### **Вывод**

**Ключевые наблюдения:**

* Прямое сравнение оказалось самым простым, но поверхностным: модели часто теряли структуру и путали элементы.
* Вопросно‑ответный метод дал более детальный анализ, но потребовал много шагов и ресурсов.
* Двухшаговое сравнение показало наиболее сбалансированный результат, позволяя чётко разделить отсутствующие, изменённые и лишние элементы.

**Оценка моделей:**

* **Qwen3** уверенно справился с задачами — держал структуру отчётов, выделял ключевые несоответствия и давал аккуратные формулировки.
* **DeepSeek‑R1** показал системность и глубину, но иногда переусложнял классификацию.
* **LLaMA‑3‑8B** и **Mistral‑7B** давали рабочие ответы, но часто ограничивались поверхностными замечаниями.

**Общее заключение:**

Наиболее качественные результаты для аналитики документов даёт **Qwen3** — модель обеспечивает точность. Среди подходов оптимальным оказался метод **двухшагового сравнения**, который позволяет фиксировать как отсутствие информации, так и изменения и лишние вставки.