# AI-ассистенты в управлении данными

SmartData
2025

Сагитов Олег| RnD-инженер в Т-Банке

Aleg Sahitau @o_qwerty789

**Работаю в направлении практик работы с Data Science**

**Более 2 лет** опыта в DG и **почти 2** в LLM-инжиниринге

# Внедрение AI в Data


Сгенерировано с помощью Mistral

# AI-ассистенты в управлении данными


С чего же начать?

# AI-ассистенты в управлении данными

Содержание

- Необходимые вводные
- Исследование, ставшее продуктом
- Техническая реализация
- Сказ о том, когда тест на проде работает
- Влияние

# Необходим ые вводные

## Есть хранилище и пользователи

Бизнес-пользователи активно строят модели данных, ETL-процессы и аналитику

# Необходимые вводные

## Есть хранилище и пользователи

Бизнес-пользователи активно строят модели данных, ETL-процессы и аналитику

## Есть Data Catalog и DQ инструменты

Пользователи закрывают свои нужды в формате Self-Service

# Необходимые вводные

## Есть хранилище и пользователи

Бизнес-пользователи активно строят модели данных, ETL-процессы и аналитику

## Есть Data Catalog и DQ инструменты

Пользователи закрывают свои нужды в формате Self-Service

## Есть центр AI-компетенций

На который можно опереться в AI (RAG, LLM serving, Evaluations, etc)

# Необходимые вводные

## Есть хранилище и пользователи

Бизнес-пользователи активно строят модели данных, ETL-процессы и аналитику

## Есть Data Catalog и DQ инструменты

Пользователи закрывают свои нужды в формате Self-Service

## Есть центр AI-компетенций

На который можно опереться в AI (RAG, LLM serving, Evaluations, etc)

## Нет идеального порядка в DG

Тогда будет веселее и интереснее!

# Начинаем исследовать!

# Необходимые вводные

# Необходимые вводные

В Q1 2025 года около **80%** пользовательских таблиц **не описаны** и их не потребляют из-за отсутствия логики в Data Catalog.

# Необходимые вводные

В Q1 2025 года около **80%** пользовательских таблиц **не описаны** и их не потребляют из-за отсутствия логики в Data Catalog.

Наличие описаний таблиц – топливо, которым надо будет уже очень скоро заправлять LLM.

Модель DWH описана на **99.9%**

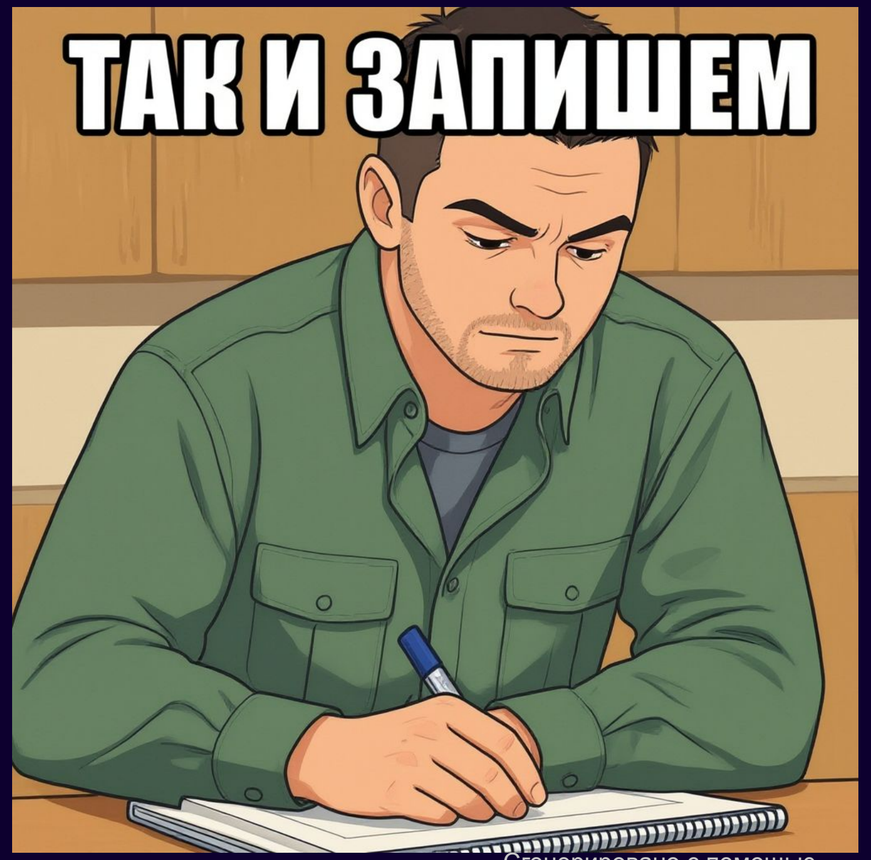
# Необходимые вводные

В Q1 2025 года около **80%** пользовательских таблиц **не описаны** и их не потребляют из-за отсутствия логики в Data Catalog.

Наличие описаний таблиц – топливо, которым надо будет уже очень скоро заправлять LLM.

Модель DWH описана на **99.9%**, мы недостаточно хорошо знаем что происходит в пользовательских слоях

# А что если помочь с описанием данных?




Сгенерировано с помощью Mistral

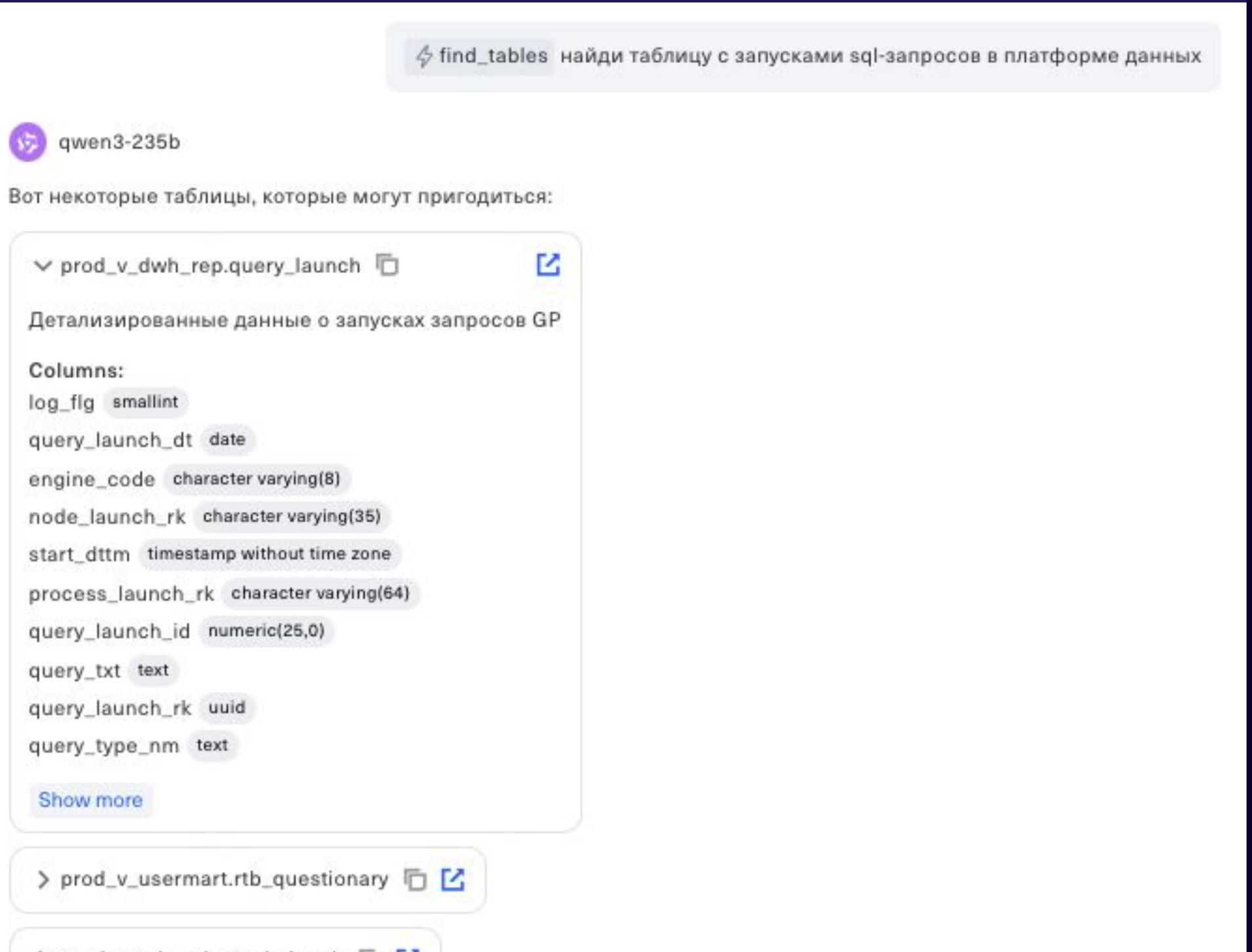
# AI-ассистенты в управлении данными

Нужно давать топливо, которым мы задрайвим другие AI-инициативы в DG

| name | description | chat_gtp_code_recommend |
|---|---|---|
| wfm_employee_rk | Постоянный ключ сотрудника | 1: 2: 3: |
| wfm_employee_id | Идентификатор оператора в WFM | 1: 2: 3: |
| hire_dt | Дата начала стажа | 1: 2:EMP 3:EMP |
| first_nm | Имя | 1:EMP 2:PER 3:PER |
| last_nm | Фамилия | 1:EMP 2:PER 3:PER |
| middle_nm | Отчество | 1:EMP 2:PER 3:PER |
| win_login | Логин в Windows | 1:EMP 2: 3:EMP |
| start_dttm | Дата начала действия записи в WFM | 1: 2: 3: |
| end_dttm | Дата окончания действия записи в WFM | 1: 2: 3: |
| timetable_txt | График работы | 1: 2: 3: |
| supervisor_nm | Супервайзер | 1:EMP 2:EMP 3: |
| department_nm | Отдел | 1: 2: 3: |
| job_title_nm | Должность | 1: 2: 3: |
| sector_nm | Сектор | 1: 2: 3: |
| siebel_login | Логин Siebel | 1:EMP 2:EMP 3:EMP |
| siebel_id | Идентификатор Siebel | 1:PID 2:PID 3:PID |
| head_sector_nm | Руководитель сектора | 1:EMP 2:EMP 3: |
| dept_head_deputy_nm | Заместитель руководителя отдела | 1:EMP 2:EMP 3: |
| employee_main_id | Идентификатор сотрудника в системе-источнике | 1: 2: 3: |
| skill_group_txt | Группа навыков | 1: 2: 3: |
| add_functionality_txt | Дополнительный функционал | 1: 2: 3: |
| management_unit_id | Идентификатор административной группы | 1: 2: 3: |
| management_unit_nm | Название административной группы | 1: 2: 3: |
| processed_dttm | Дата и время обработки записи ETL-процессом | 1: 2: 3: |

## Data Security Advisor

Рекомендательная система на базе описаний по разметке ЧД в таблицах

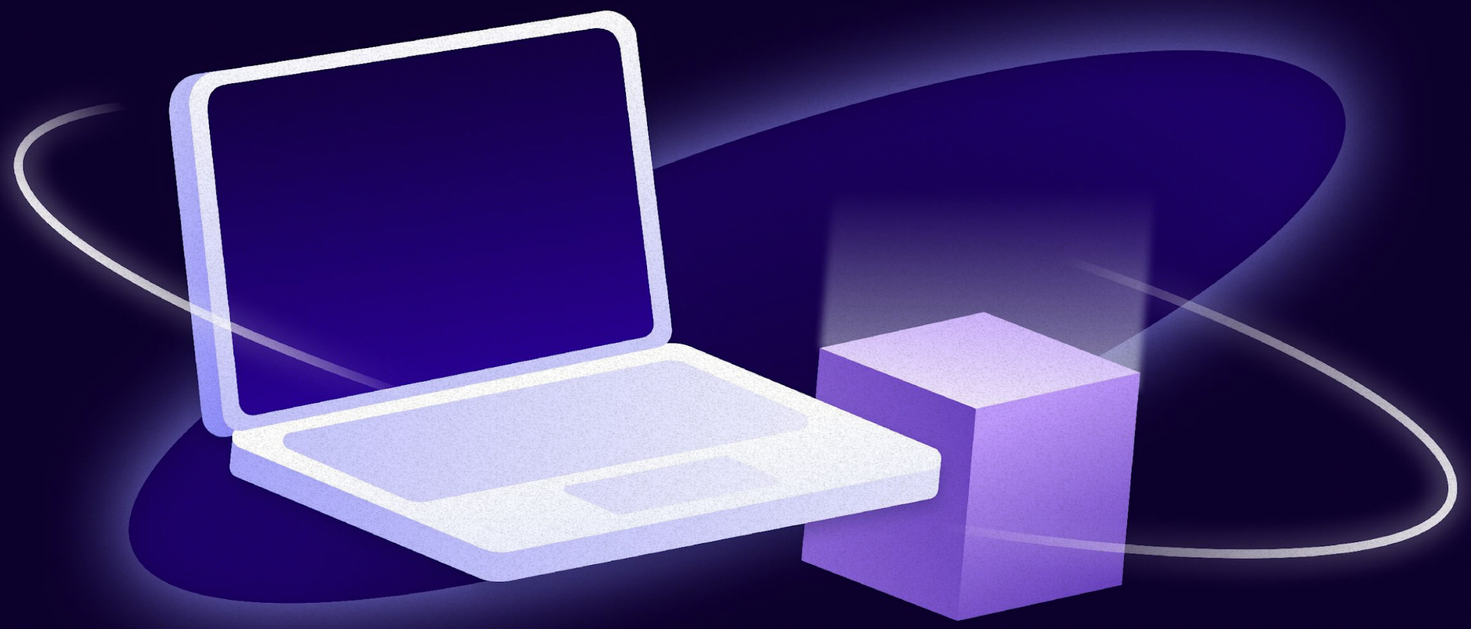



# AI Table Search

Семантический поиск по таблицам Data Platform

# Исследование, ставшее продуктом

Или как RnD делает продукты из гипотез

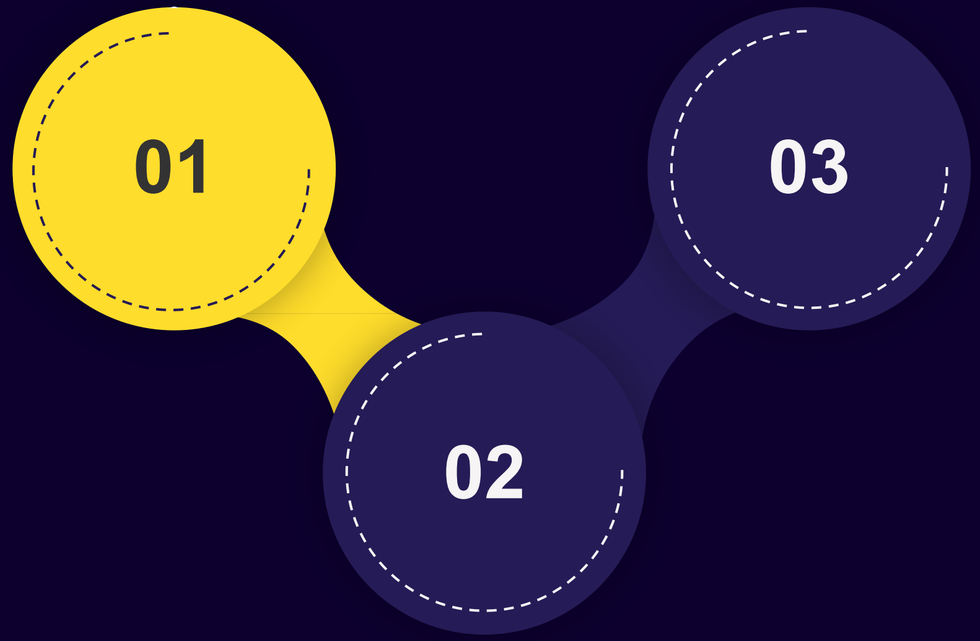


Исследование, ставшее продуктом

# С чего начали?

## Определили точку старта

Провели эксперименты о применимости в пайплайне различных данных



Исследование, ставшее продуктом

# С чего начали?

## Определили точку старта

Провели эксперименты о применимости в пайплайне различных данных


## Собирали ОС

CustDev указал, что часто встречаются близкие формулировки, но не в терминологии

Исследование, ставшее продуктом

# С чего начали?

## Определили точку старта

Провели эксперименты о применимости в пайплайне различных данных


## Собирали ОС

CustDev указал, что часто встречаются близкие формулировки, но не в терминологии

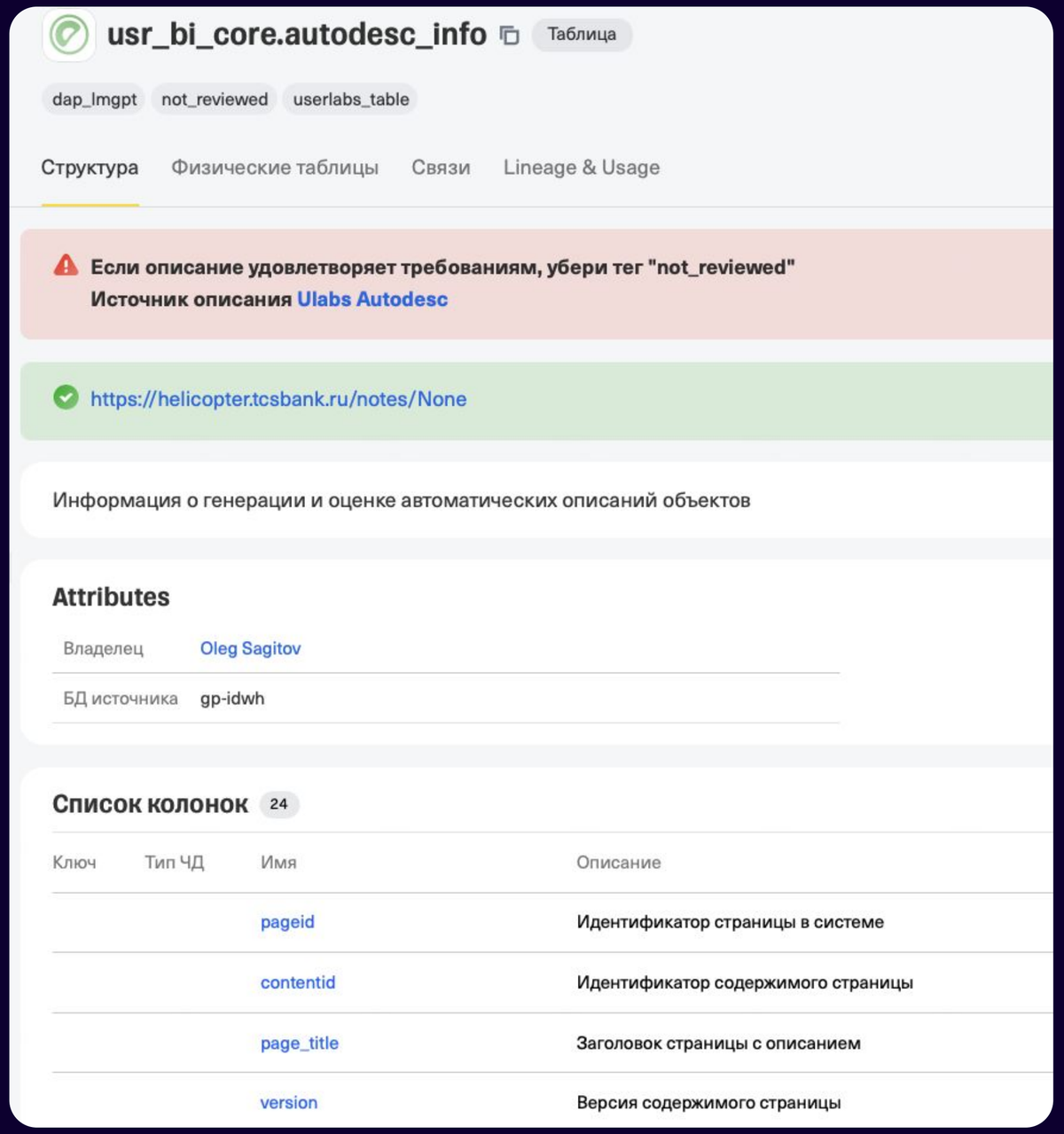
## Пошли в прод

Реализовав все фич-реквесты нет смысла медлить…

# Инструмент для генерации описаний таблиц

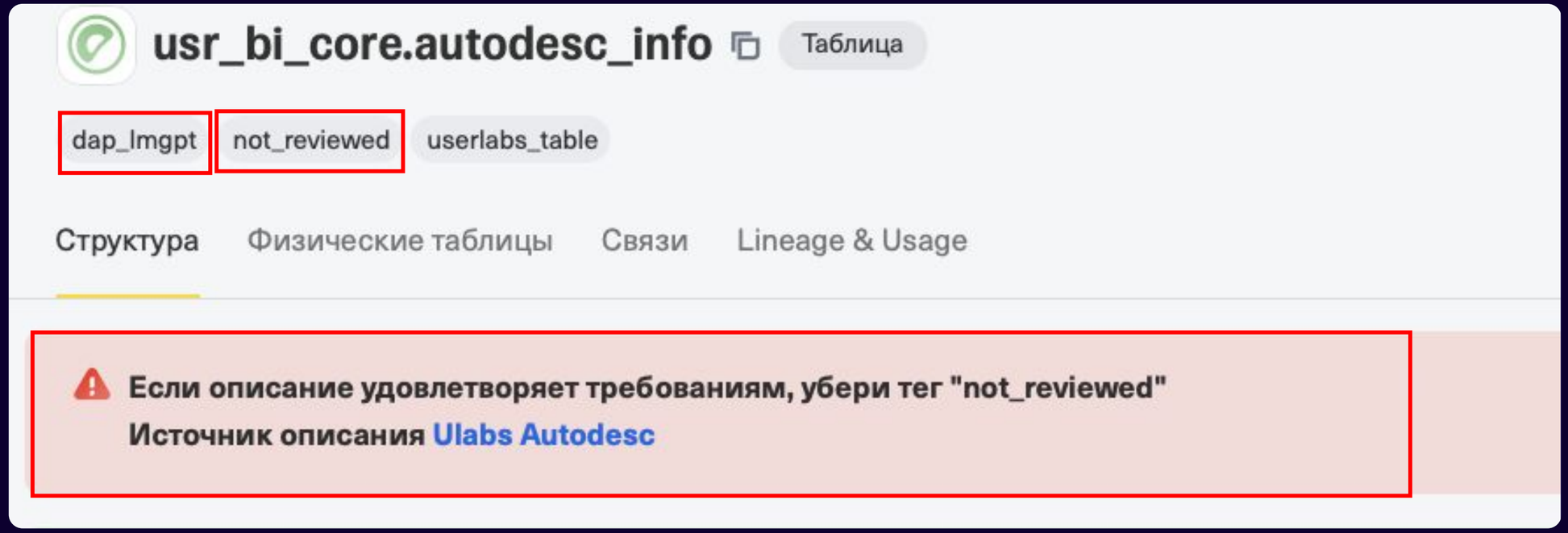
autodes
c

Исследование, ставшее продуктом



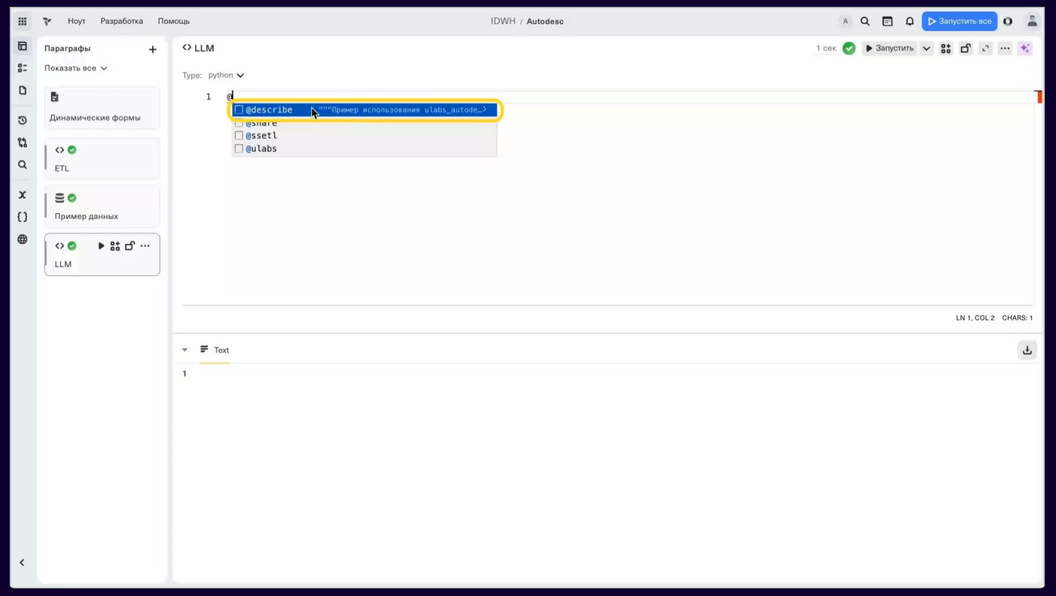
# Инструмент для генерации описаний таблиц autodesc

Исследование, ставшее продуктом



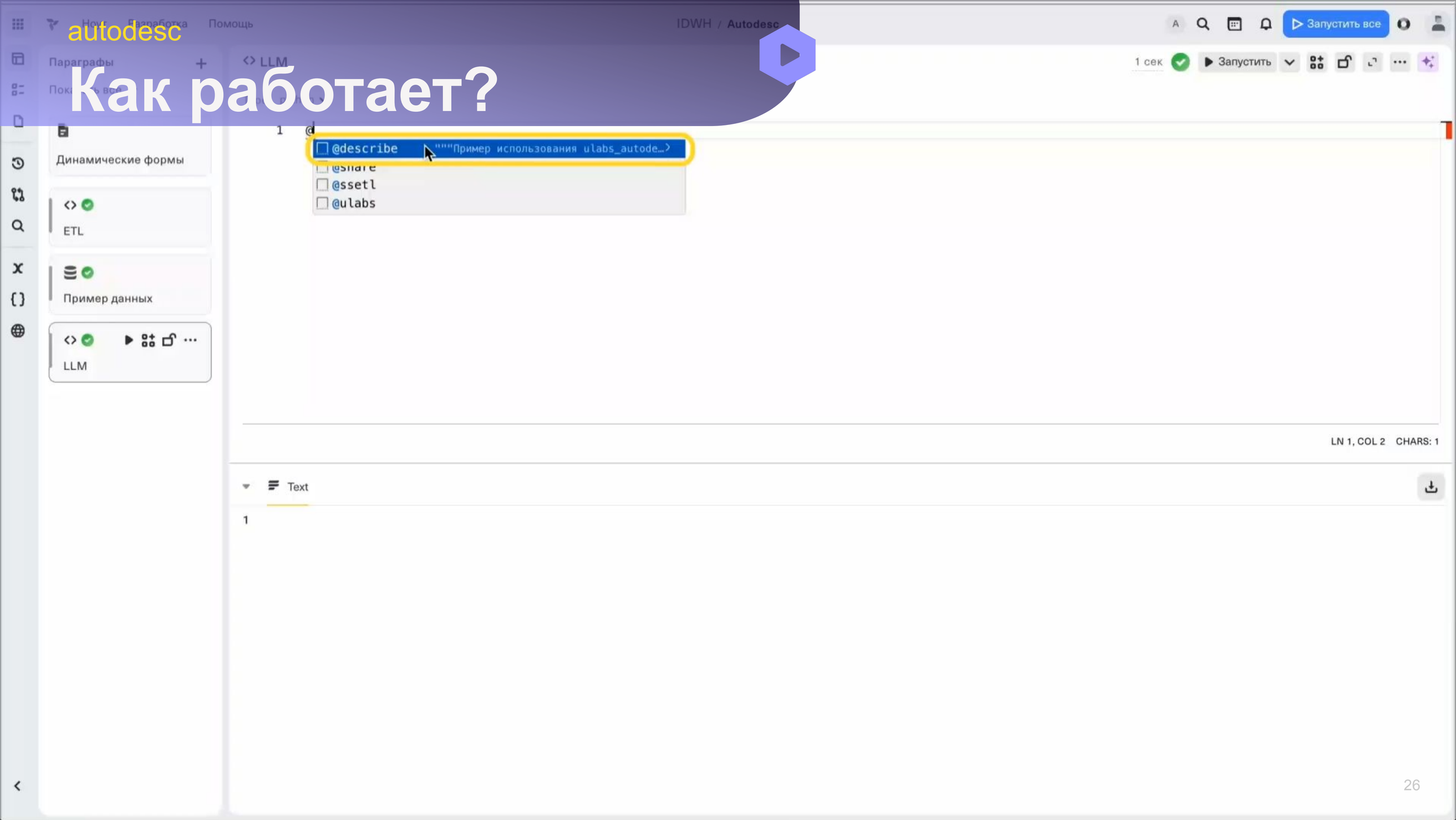
# autodesc

Исследование, ставшее продуктом



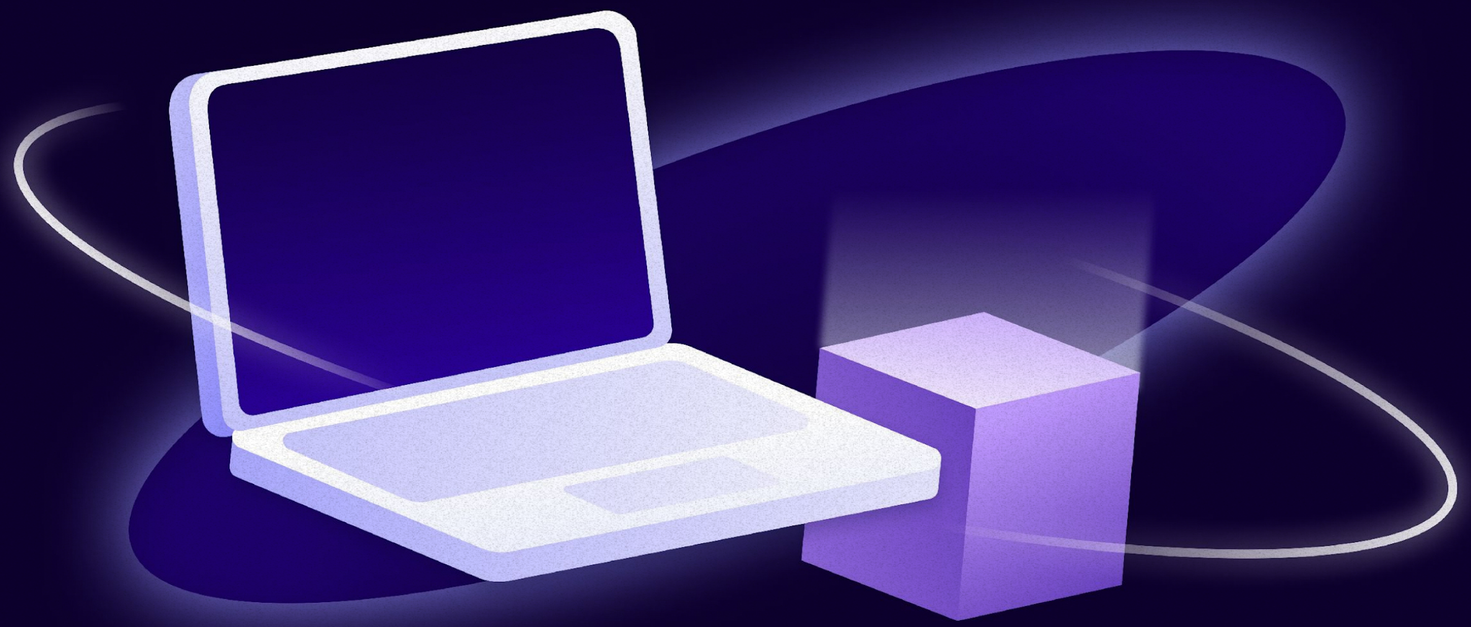

autodesc

# Как работает?



# Техническая реализация

autodesc



# autodesc

**Как новый подход повлиял на генерацию?**

Было

| | |
|---|---|
| party_rk | Ключ партии |

Стало

| | |
|---|---|
| party_rk | Уникальный ключ клиента в DWH |

# autodesc

## Пример работы механизмов

**Стандарты проектирования:**

- {"suffix": "_rk", "text": "Постоянный ключ экземпляра сущности в DWH", "category": "Идентификаторы", "examples": "account_rk, party_rk, auto_payment_template_rk"}
- {"entity_name": "party", "entity_desc": "Участник процесса"}

**Golden RAG:**

- {"column_nm": "party_rk","column_description": "Постоянный ключ участника"},

**Глоссарий:**

- {"term": "Клиент","content": "Клиент банка или других продуктов холдинга \"Т\" (инвестиции, страхование, мобайл).","synonyms":"Пользователь приложения, Участник, Покупатель"}

- За счет контекста по стандартам LLM начинает понимать, что речь о уникальном ключе в DWH
- За счет глоссария – обретает гибкость в контексте, из-за чего в генерации появилось уточнение «Клиент»

# autodesc

Пример работы механизмов

**Стандарты проектирования:**

- {"suffix": "_rk", "text": "Постоянный **ключ** экземпляра сущности **в DWH**", "category": "Идентификаторы", "examples": "account_rk, **party_rk**, auto_payment_template_rk"}
- {"entity_name": "party", "entity_desc": "**Участник процесса**"}

**Golden RAG:**

- {"column_nm": "party_rk","column_description": "Постоянный **ключ участника**"},

**Глоссарий:**

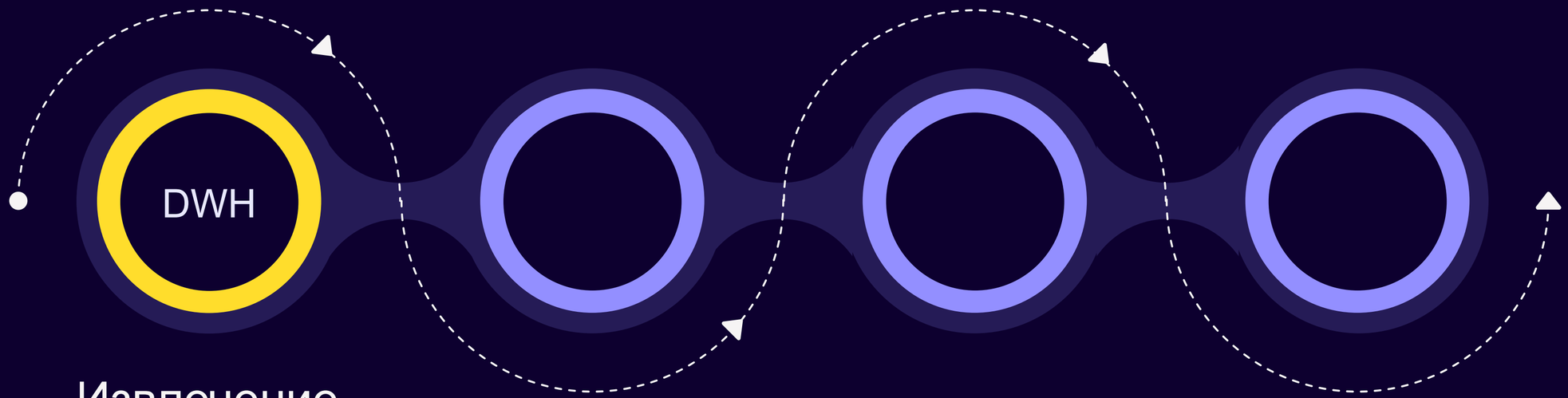
- {"term": "**Клиент**","content": "Клиент банка или других продуктов холдинга \"Т\" (инвестиции, страхование, мобайл).","synonyms":"Пользователь приложения, Участник, Покупатель"}

autodesc

# Механизм описания

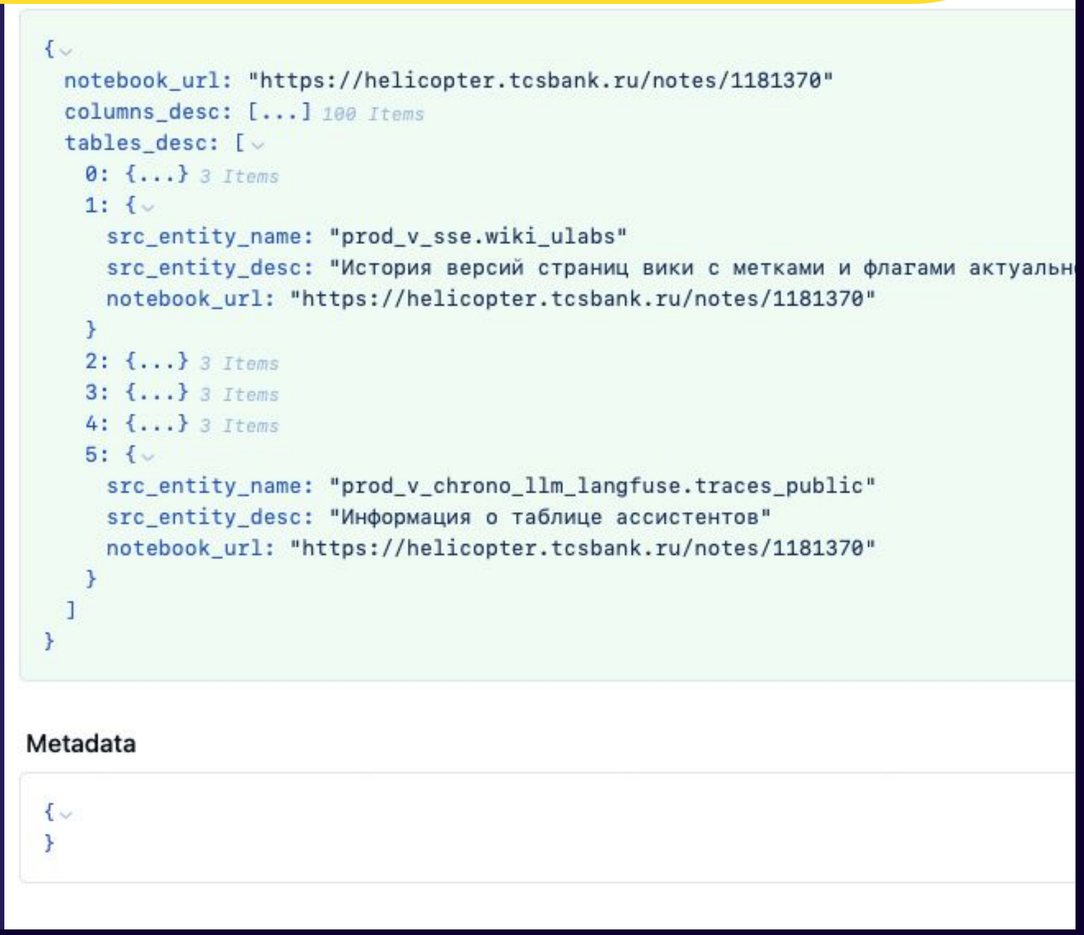
autodesc

# Механизм описания




Извлечение данных
Lineage + Sample

# Механизм описания

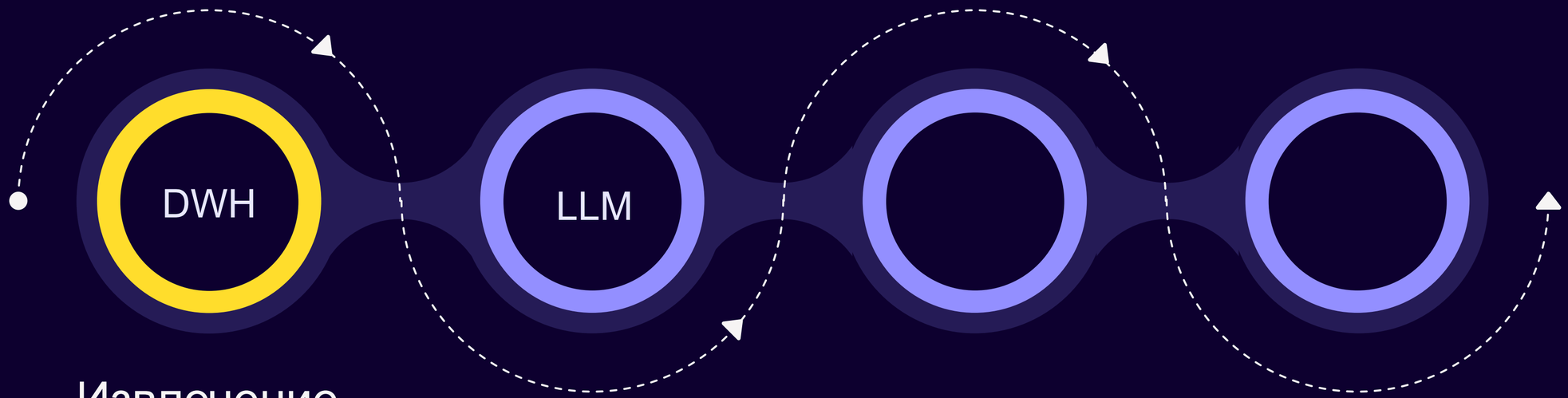



Lineage + Sample

Считывая Lineage и содержимое таблиц – мы обеспечиваем понимание происхождения и состояния данных.

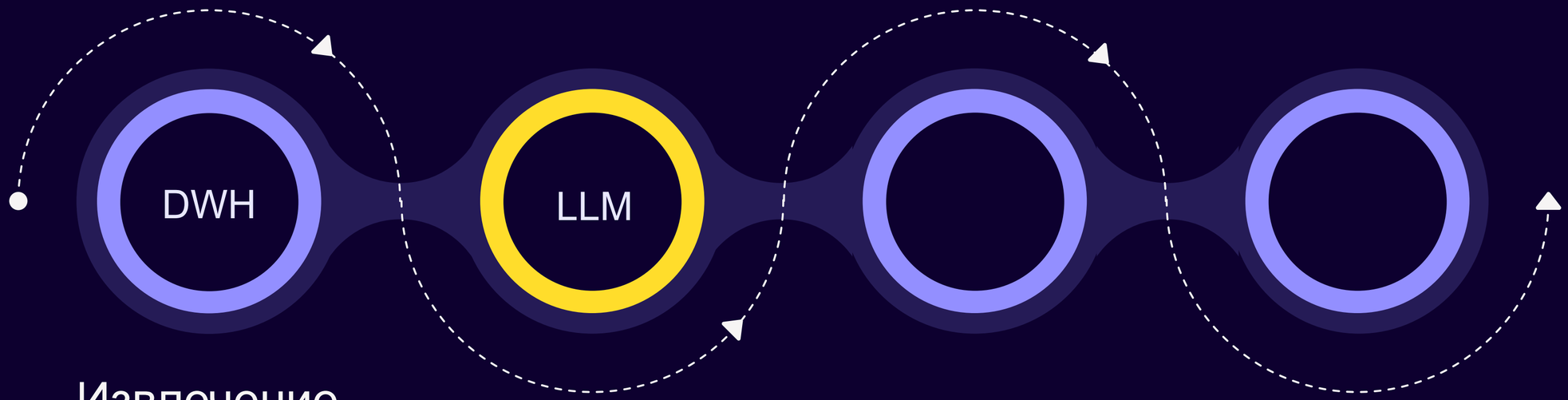
autodesc

# Механизм описания



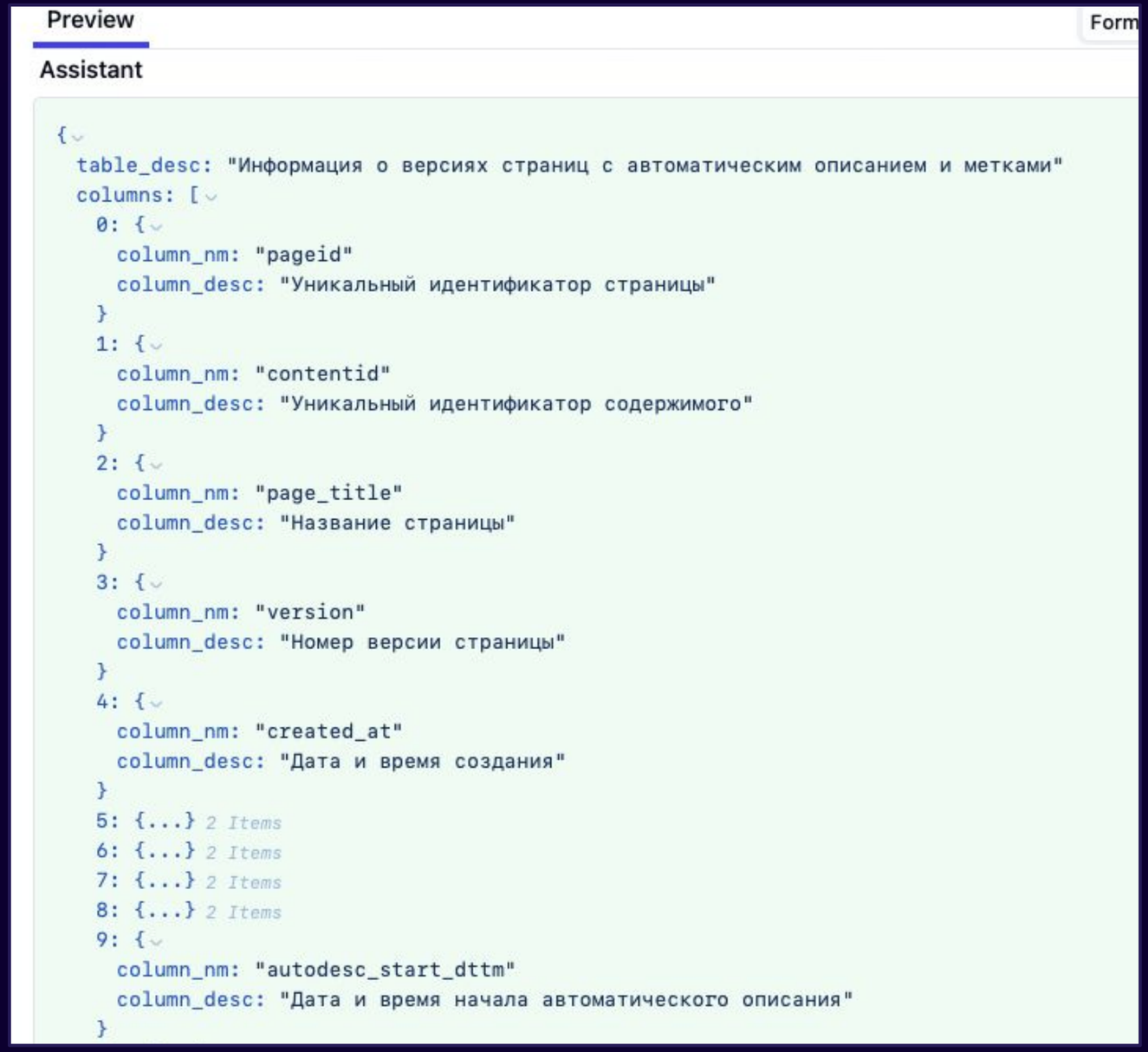

Извлечение данных
Lineage + Sample

# Механизм описания




Извлечение данных
Lineage + Sample

Генерация
Lineage + Sample




autodesc

# Генерация

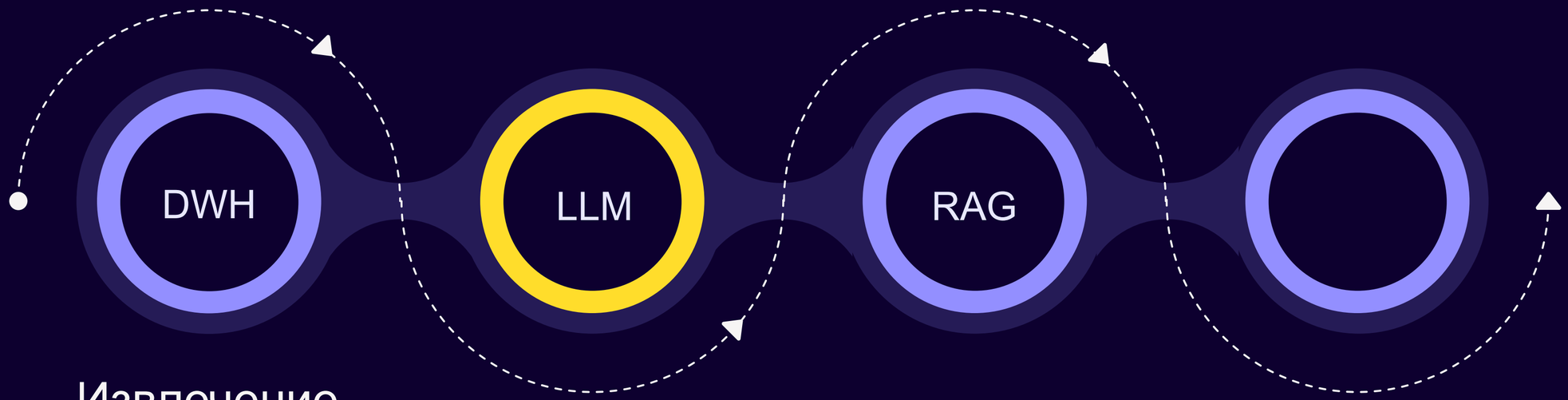
Первый наш этап генерации описания, где с глубоким пониманием рождается описание.

Сэмпл данных позволяет не путаться при ломанном именовании столбцов

Lineage позволяет точнее выявлять взаимосвязь между данными

autodesc

# Механизм описания

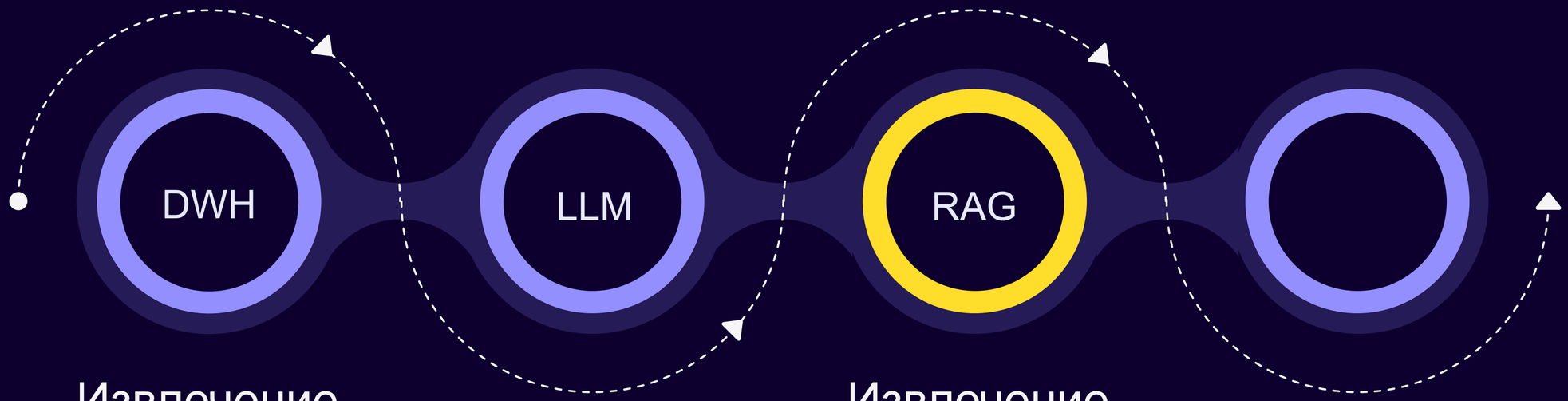



Извлечение данных
Lineage + Sample

Генерация
Lineage + Sample

autodesc

# Механизм описания




**Извлечение данных**
Lineage + Sample

**Генерация**
Lineage + Sample

**Извлечение данных**
Стандарты +
Goldens +
Глоссарий

# Как логически формируем Golden-ы?

Есть потребление

Топ-3 самых потребляемых таблицы по схеме

# Как логически формируем Golden-ы?

## Есть потребление

Топ-3 самых потребляемых таблицы по схеме

## Эвристики

Более 5 символов и двух слов в описании таблицы.

Более 3 символов и одного слова в описании столбцов.

# Как логически формируем Golden-ы?

## Есть потребление

Топ-3 самых потребляемых таблицы по схеме

## Эвристики

Более 5 символов и двух слов в описании таблицы.

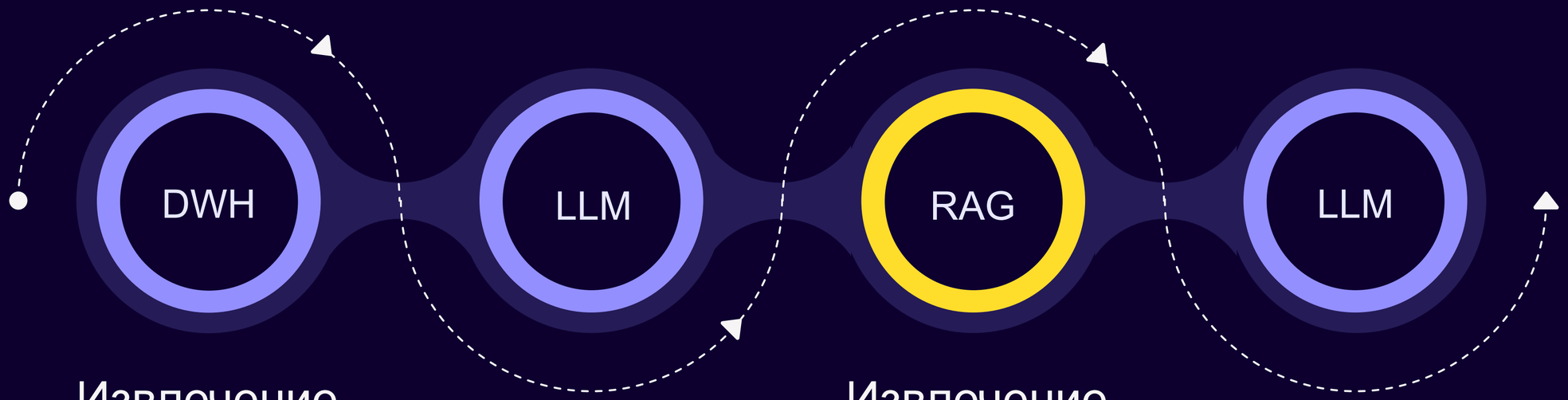
Более 3 символов и одного слова в описании столбцов.

## Валидация бизнеса

Бизнес явно выделил ценность объекта и связал его с терминами

autodesc

# Механизм описания



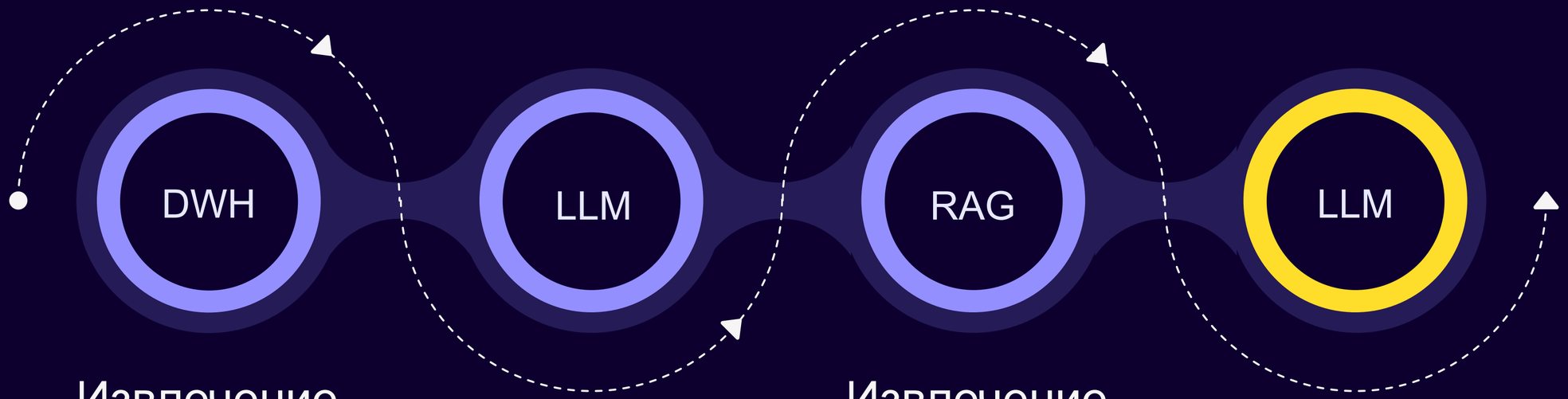

**Извлечение данных**
Lineage + Sample

**Генерация**
Lineage + Sample

**Извлечение данных**
Стандарты +
Goldens +
Глоссарий

autodesc

# Механизм описания



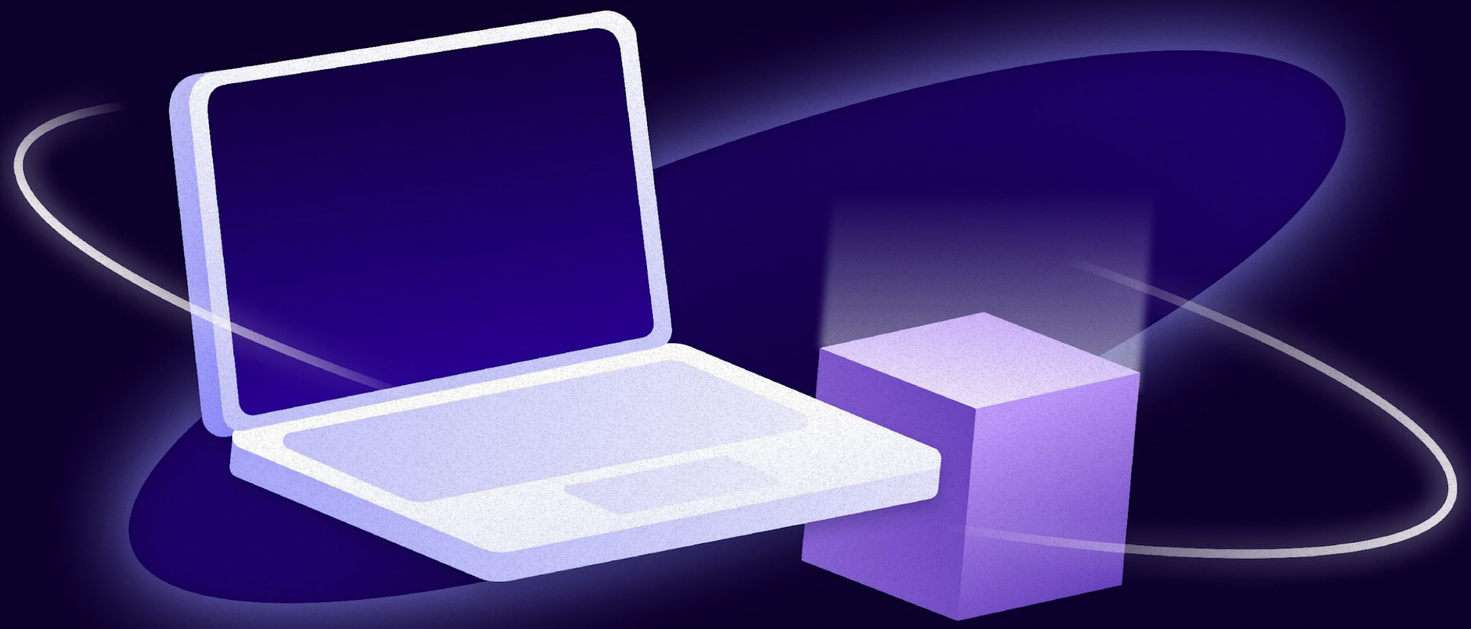

**Извлечение данных**
Lineage + Sample

**Генерация**
Lineage + Sample

**Извлечение данных**
Стандарты +
Goldens +
Глоссарий

**Генерация**
Описание +
Стандарты +
Goldens +
Глоссарий

# Сказ о том, когда тест на проде работает



# Тест на проде работает?

```
42 +     # act
43 +     with pytest.raises(ConfigError, match=f"Path {path} leads to a not yaml file"):
44 +         build_features_static_data_from_yaml_config(path)
45 +
46 +
47 + def test_build_config_map__config_file_is_empty(prepared_features_config: TConfigMap):
48 +     # act
49 +     with pytest.raises(
50 +         pydantic.ValidationError,
51 +         match="Dictionary should have at least 1 item",
```

**Review Bot** @review-bot · 1 month ago

Developer

Рандомное текстовое содержимое новообразованного кода, предназначенного для тестирования яиц от OEI вашего дядьки, которого ты застукал в подвале с Брежневым.完全なる混沌.

🤖 Generated by Gitlab Review tool

👍 24

# Тест на проде

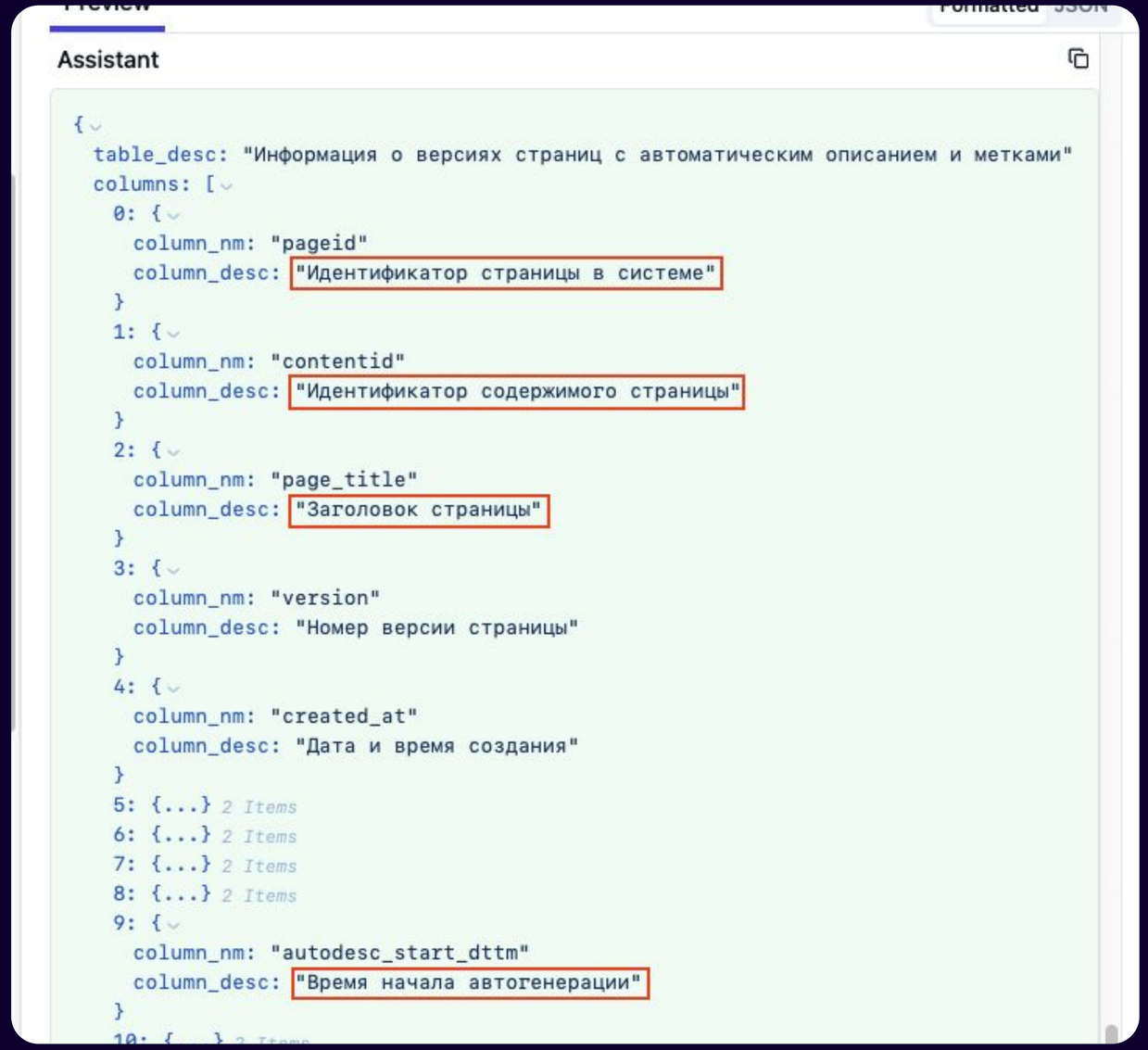
Preview Format

Assistant


```
{
  table_desc: "Информация о версиях страниц с автоматическим описанием и метками"
  columns: [
    0: {
      column_nm: "pageid"
      column_desc: "Уникальный идентификатор страницы"
    }
    1: {
      column_nm: "contentid"
      column_desc: "Уникальный идентификатор содержимого"
    }
    2: {
      column_nm: "page_title"
      column_desc: "Название страницы"
    }
    3: {
      column_nm: "version"
      column_desc: "Номер версии страницы"
    }
    4: {
      column_nm: "created_at"
      column_desc: "Дата и время создания"
    }
    5: {...} 2 Items
    6: {...} 2 Items
    7: {...} 2 Items
    8: {...} 2 Items
    9: {
      column_nm: "autodesc_start_dttm"
      column_desc: "Дата и время начала автоматического описания"
    }
    10: {...} 2 Items
```

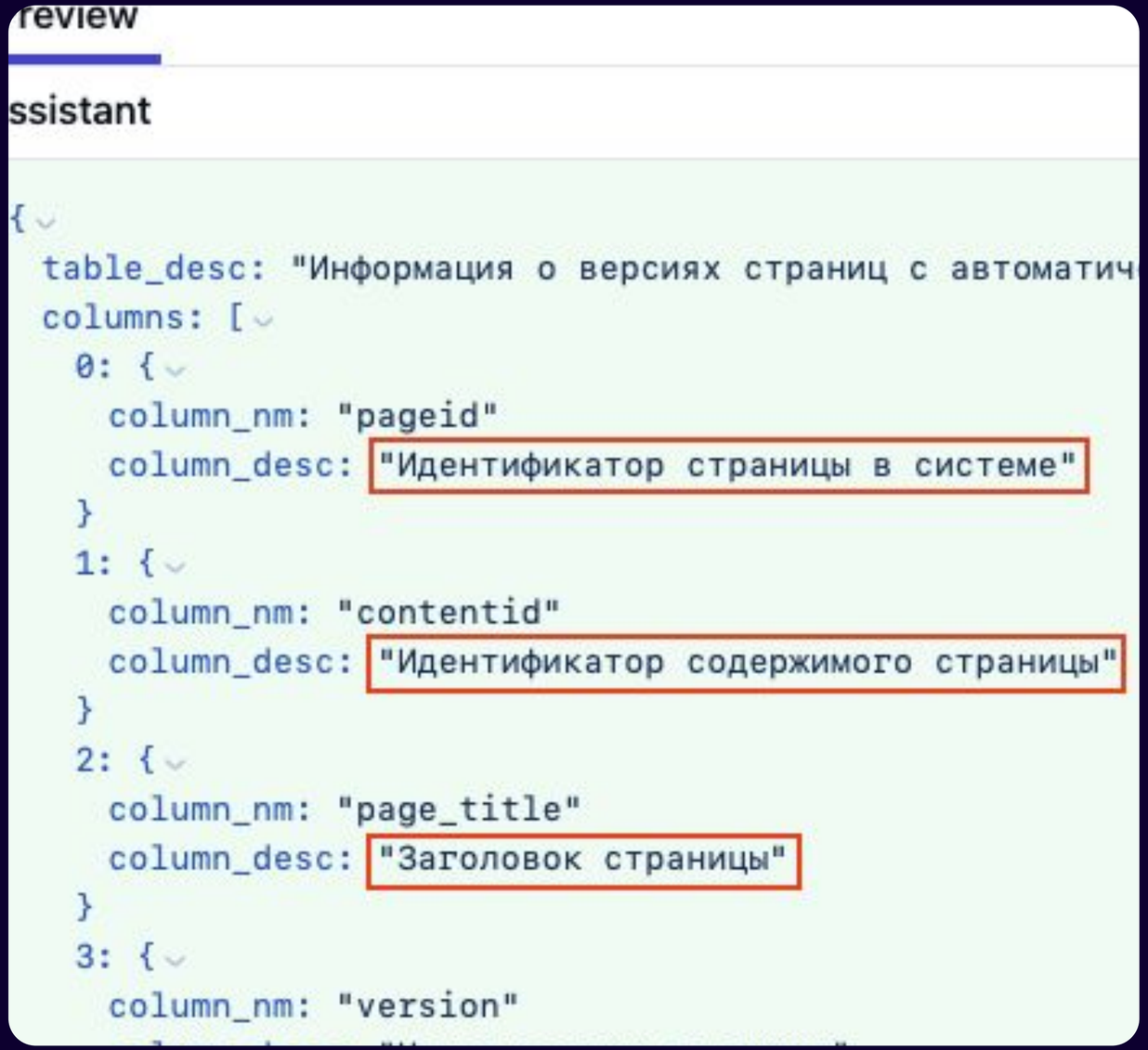

- **Неточное** описание **pageid**
- **Неконкретное** описание **contentid**
- **Избыточно** развернутое описание **autodesc_start_dttm**

# Тест на проде



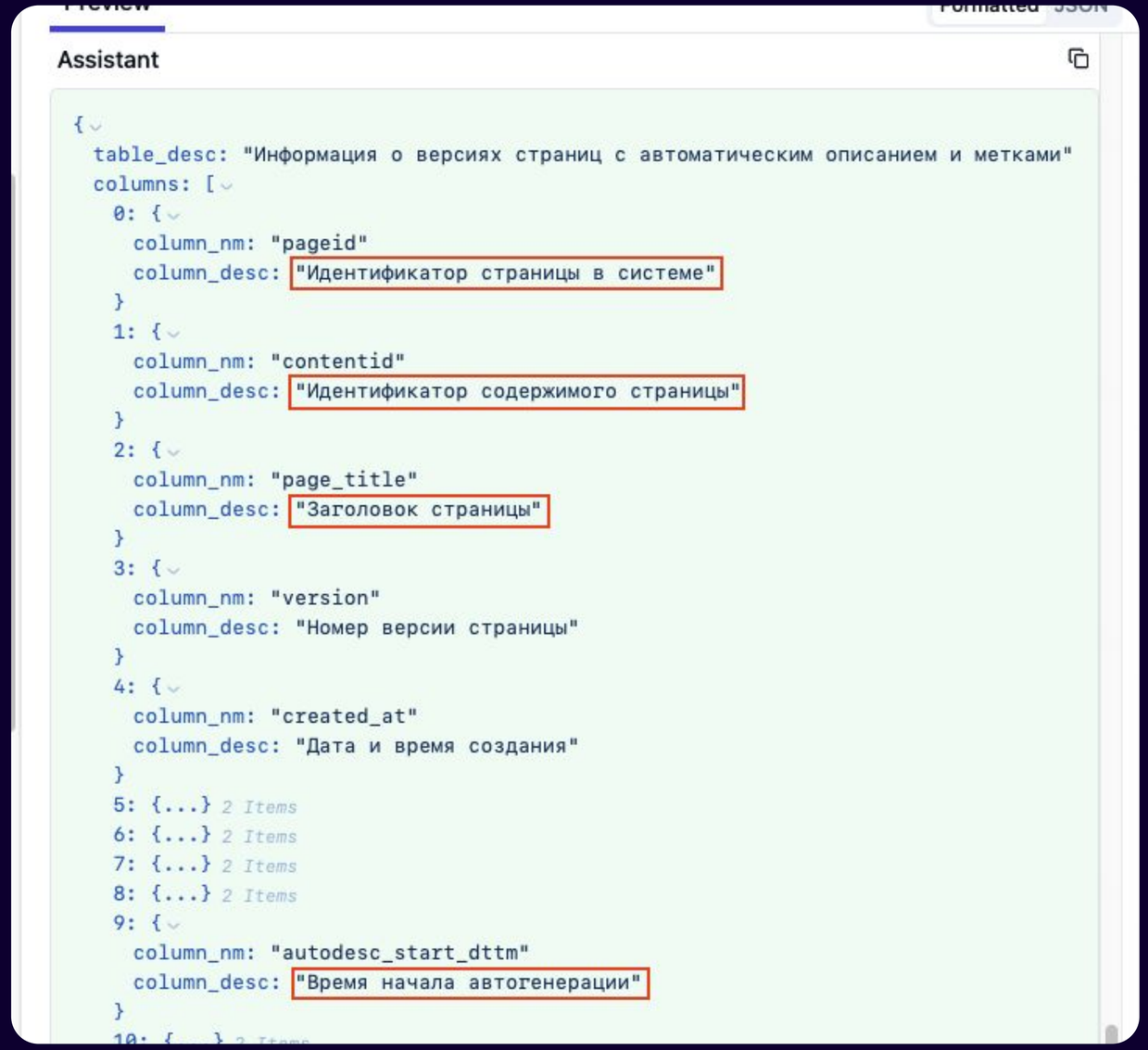

- **Уточнение pageid** как идентификатора в системе
- **Конкретизация contentid** до «страницы»
- **Сокращение** и применение терминологии **autodesc_start_dttm**

# Тест на проде




- **Уточнение pageid** как идентификатора в системе
- **Конкретизация contentid** до «страницы»
- **Сокращение** и применение терминологии **autodesc_start_dttm**

# Тест на проде




- **Уточнение pageid** как идентификатора в системе
- **Конкретизация contentid** до «страницы»
- **Сокращение** и применение терминологии **autodesc_start_dttm**

# Online Eval

## Тест на проде

0%

### Cosine Simularity

Медиана матрицы косинусного сходства для N генераций.

0%

### ContextPrecision@k

Классика из фреймворка RAGAS.

Метрикой отвечаем на вопрос: «Нашлось ли что-то полезное?»

0%

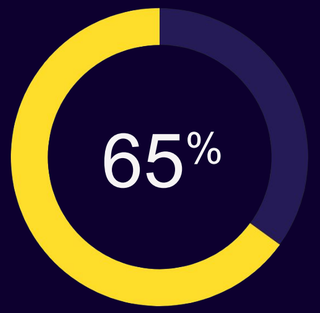
### Answer Faithfulness

Немного трансформированная RAGAS-метрика.

На основе input и контекста из RAG формируем другой LLM утверждения, затем оцениваем ответ с утверждениями.

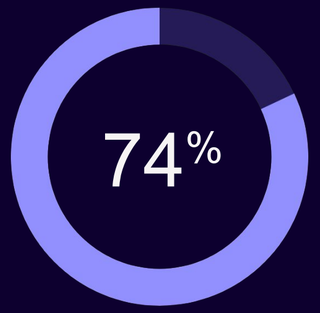
# Online Eval

## Тест на проде




### Cosine Simularity

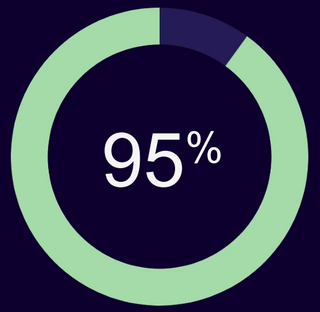
Медиана матрицы косинусного сходства для N генераций.




### ContextPrecision@k

Классика из фреймворка RAGAS.

Метрикой отвечаем на вопрос: «Нашлось ли что-то полезное?»




### Answer Faithfulness

Немного трансформированная RAGAS-метрика.

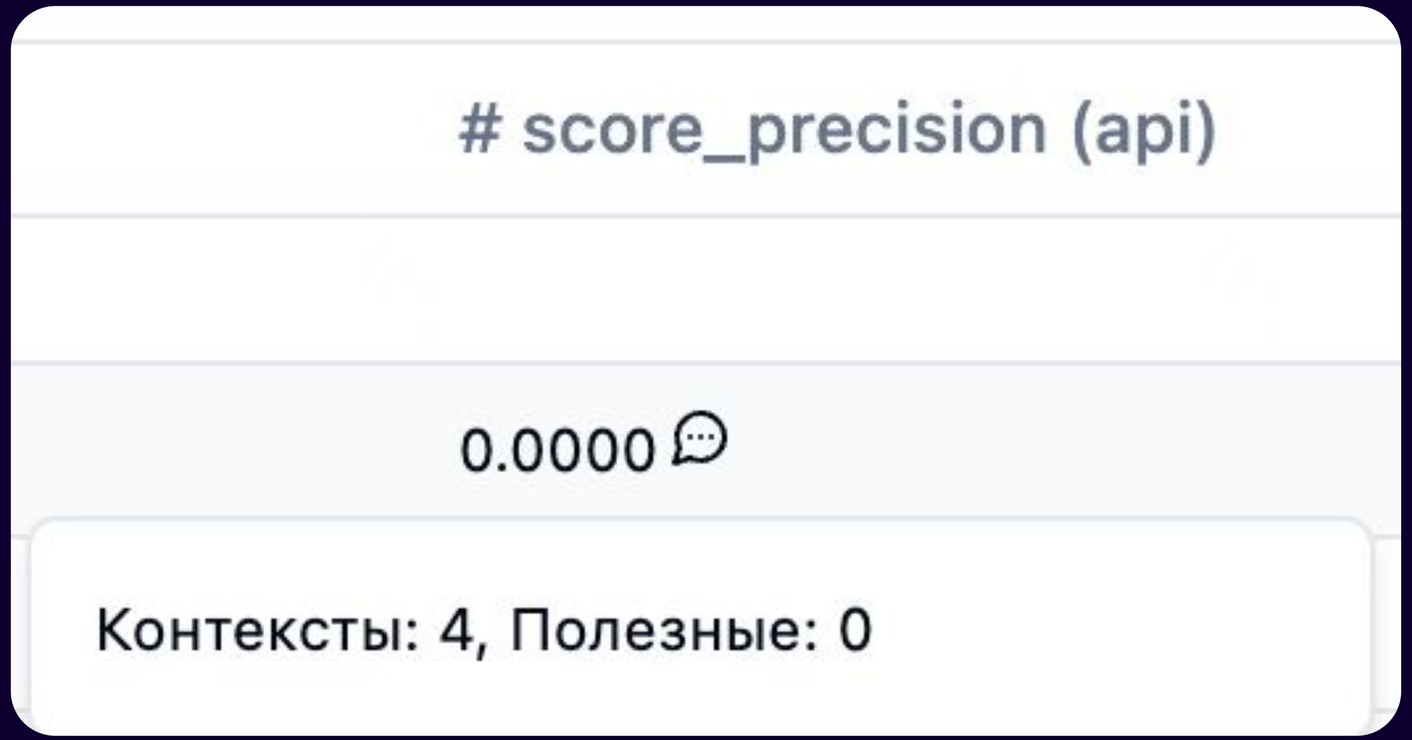
На основе input и контекста из RAG формируем другой LLM утверждения, затем оцениваем ответ с утверждениями.


# Наблюдаемость

Позволяет без лишних приседаний

увидеть резкое падение метрик день ото

дня

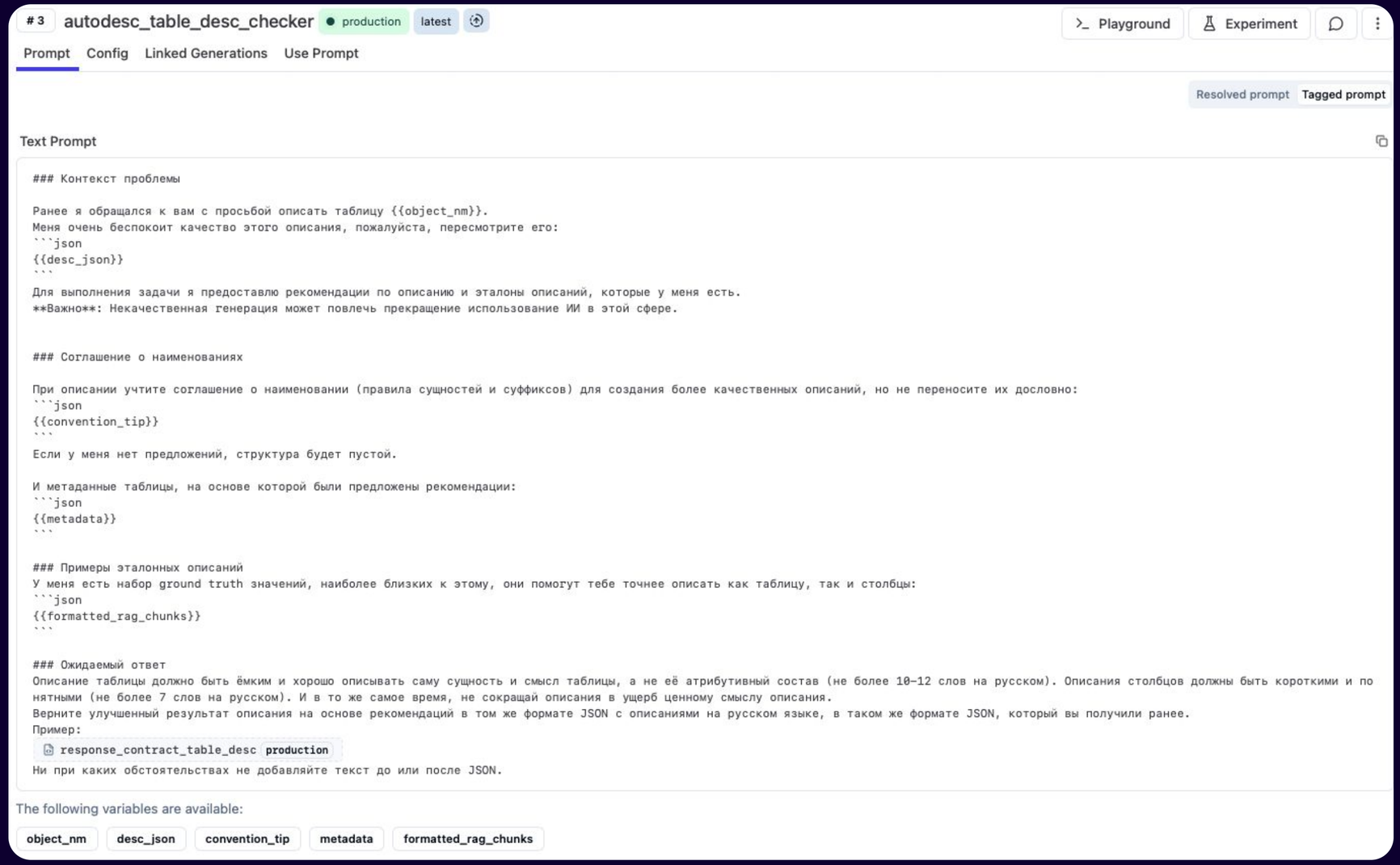



## Прозрачность

При падении, есть возможность углубиться в детали и провести диагностику глубже

# Управляемость

## Prompt-Management

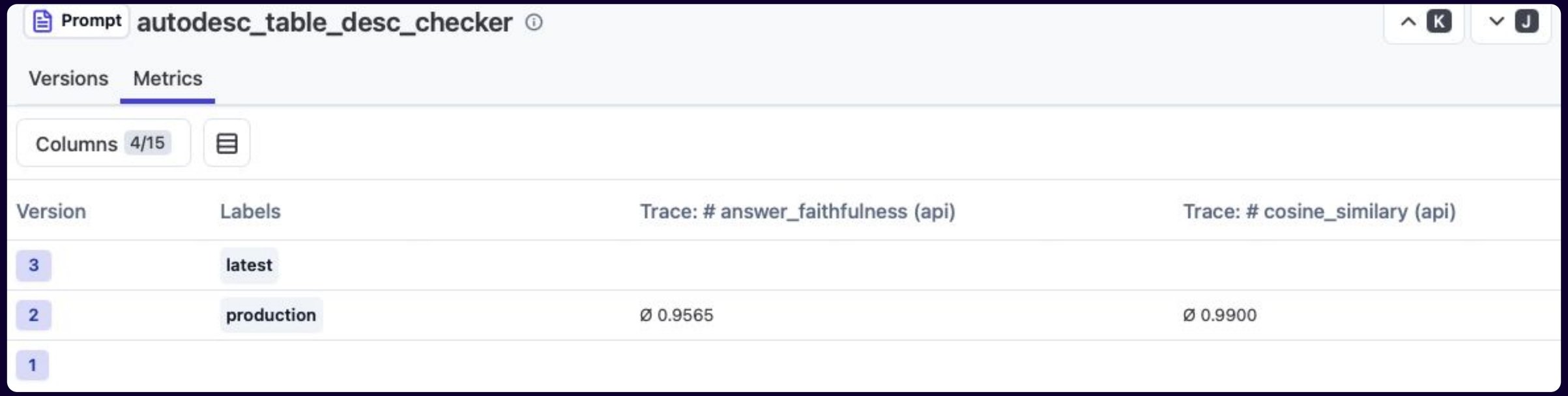


# Управляемость

Модульность промптов

```
### Ожидаемый ответ
Описание таблицы должно быть ёмким и хорошо описывать саму сущность и смысл таблицы, а не её атрибутивный сос
нятными (не более 7 слов на русском). И в то же самое время, не сокращай описания в ущерб ценному смыслу опис
Верните улучшенный результат описания на основе рекомендаций в том же формате JSON с описаниями на русском яз
Пример:
  response_contract_table_desc production
Ни при каких обстоятельствах не добавляйте текст до или после JSON.
```

# Eval + Prompt-Management



Технические метрики

# autodesc

## Latency

~50sec

На основе нескольких CustDev и собственных замеров – описать таблицу руками дело ~12 минут.

Технические метрики

# autodesc

## Latency

~50sec

На основе нескольких CustDev и собственных замеров – описать таблицу руками дело ~12 минут.

## Cost

~0.07$

Или 20k токенов на описание одной таблицы размером в диапазоне от 10 до 20 столбцов.

# Влияние

Метрики adoption

# Влияние

CSAT

4.4

На основе опроса
~150 респондентов.

Метрики adoption

# Влияние

## CSAT

4.4

На основе опроса более чем 150 респондентов.

## WAU

59

%

От всех аналитиков создающих описания в DWH (210+)

Метрики adoption

# Влияние

| CSAT | WAU | Доля AI-контента |
| --- | --- | --- |
| 4.4 | 59 % | 42 % |
| На основе опроса более чем 150 респондентов. | От всех аналитиков создающих описания в DWH (210+) | Уже почти половина пользовательских-таблиц описаны с помощью AI. |

# Влияние на пользователей

Как повлияло на взаимодействие с платформой?


Применение для дашбордов и песочниц

- Автоматизация покрытия пользовательских доменов описаниями
- Облегчение и развитие поиска по аналитическим артефактам

# Влияние на пользователей

Как повлияло на взаимодействие с платформой?

**ANALYTICS**

Применение для дашбордов и песочниц

- Автоматизация покрытия пользовательских доменов описаниями
- Облегчение и развитие поиска по аналитическим артефактам

**SSETL**

Встраивание в пайплайн ETL-разработки:

- Ускорение автоматизированного ревью
- Упрощение актуализации при доработках

# Влияние на пользователей

Как повлияло на взаимодействие с платформой?

**ANALYTICS**

Применение для дашбордов и песочниц

- Автоматизация покрытия пользовательских доменов описаниями
- Облегчение и развитие поиска по аналитическим артефактам

**SSETL**

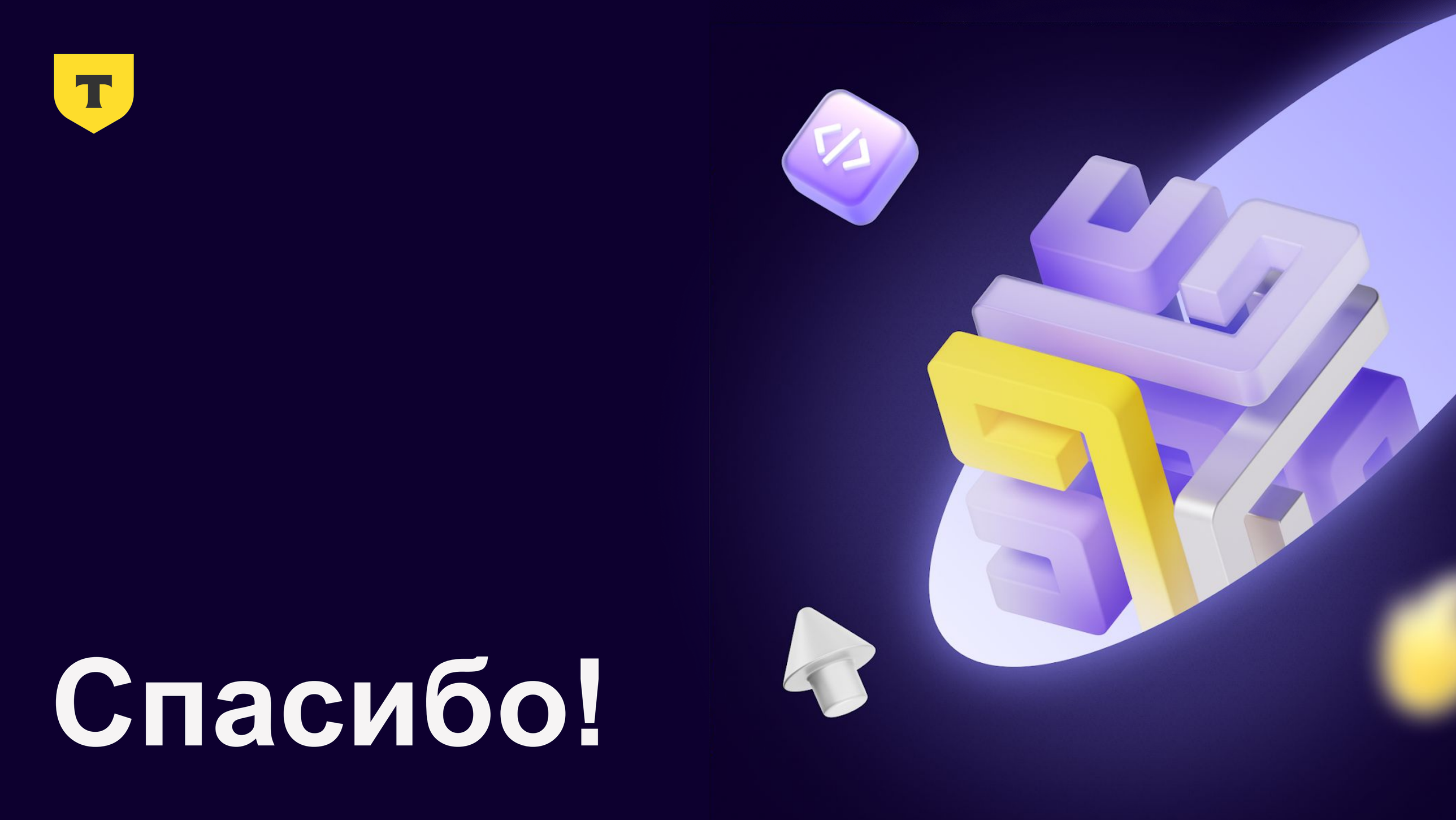
Встраивание в пайплайн ETL-разработки:

- Ускорение автоматизированного ревью
- Упрощение актуализации при доработках

**ML**

Инициализация фич для каталогизации:

- Оптимизация взаимодействия с платформой
- Выделение аналитической ценности



# Спасибо!