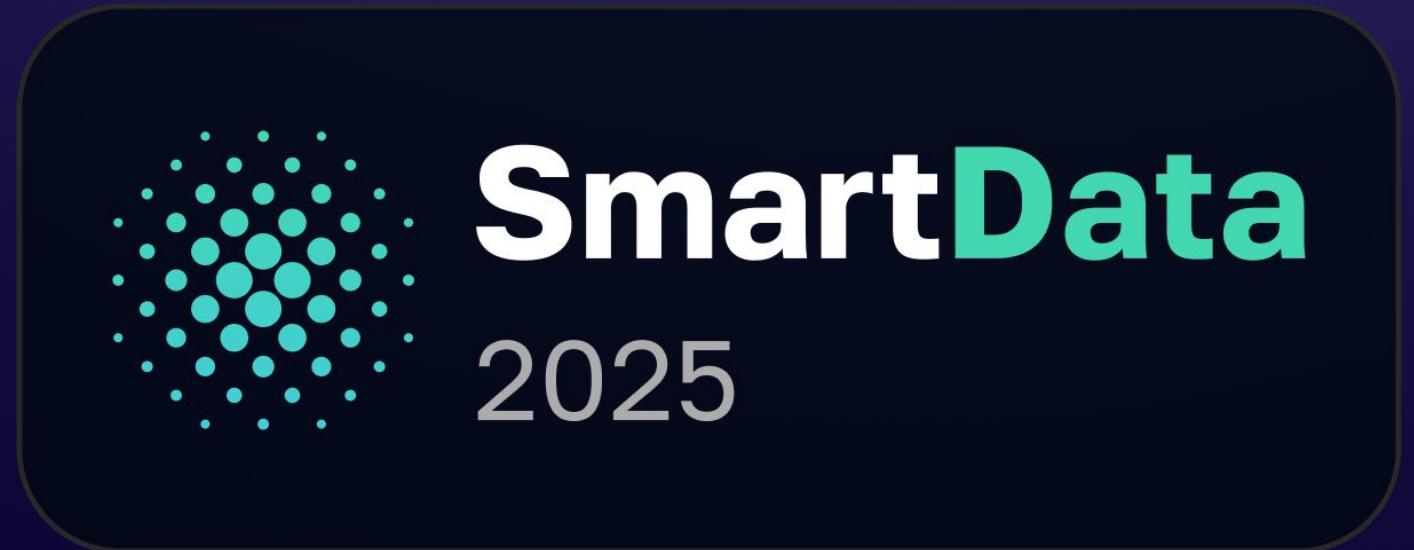


AI-ассистенты в управлении данными





Работаю в
направлении
практик
работы с



Деял в ² имеет
опыта в DG и
почти 2 в LLM-
инжиниринге

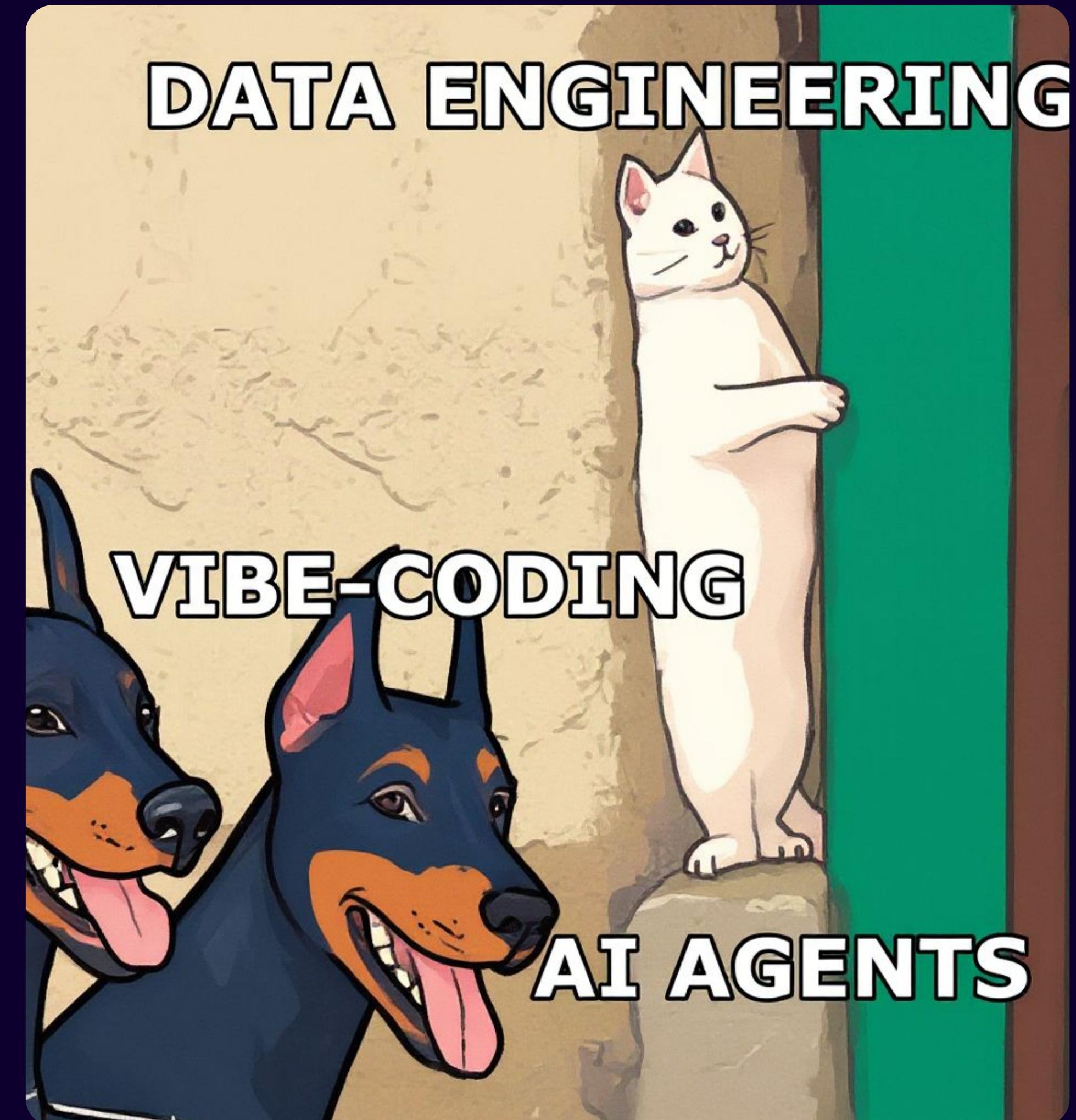
Сагитов Олег | RnD-инженер в Т-

[Баике Sagitau](#)



[@o_qwerty789](#)

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

Необходимые вводные



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

Необходимые вводные



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



Есть Data Catalog и DQ
Использование выравнивают свои нужды в
формате Self-Service

Необходимые вводные



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



Есть Data Catalog и DQ
Использование выстраивают 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
Бизнес-пользователи активно строят
Новые модели данных, ETL-процессы и аналитику



Есть Data Catalog и DQ
Использование выстраивают 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
ДГа 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



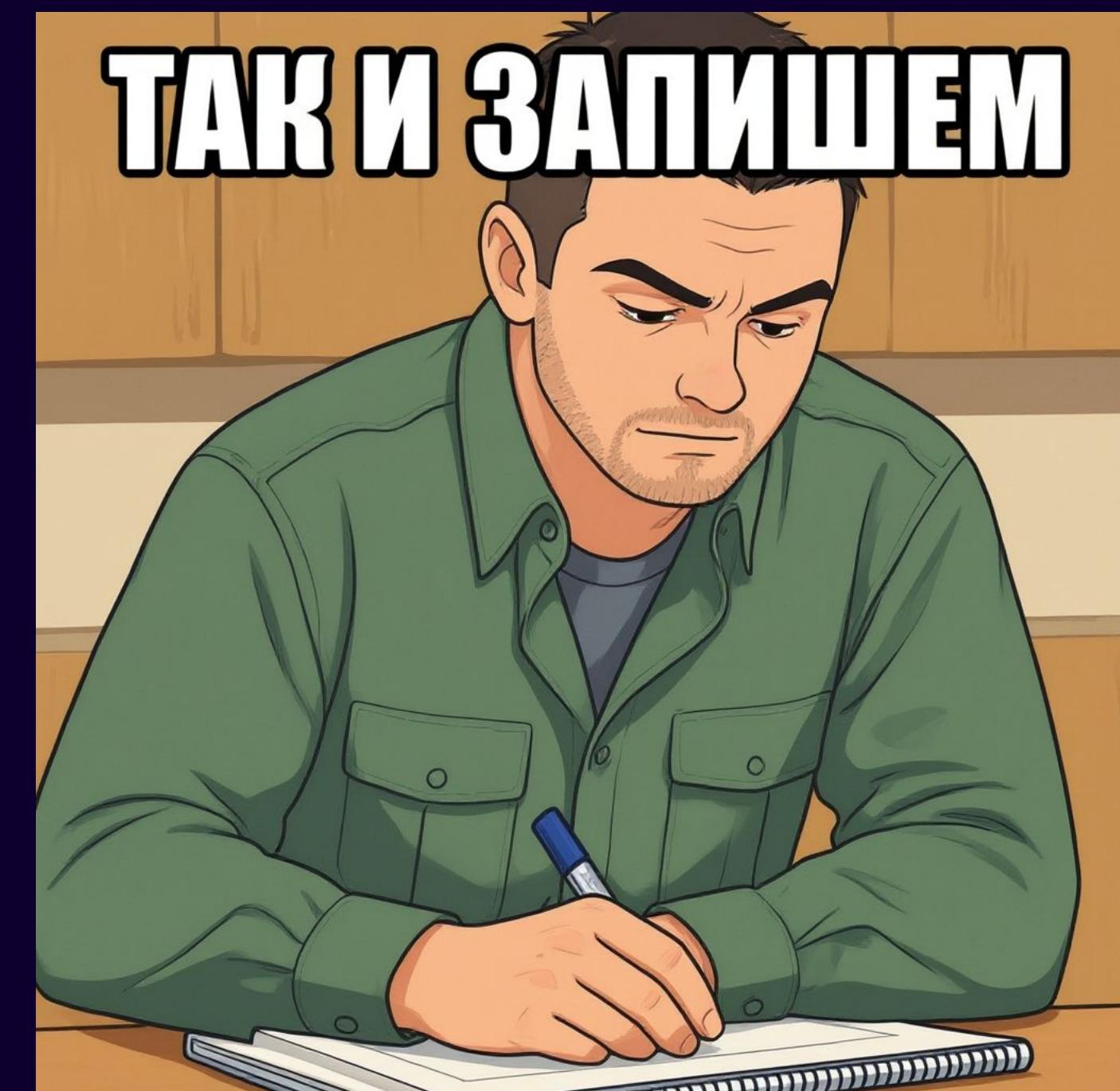
В Q1 2025 года около **80%** пользовательских таблиц **не описаны** и их не потребляют из-за отсутствия логики в Data Catalog.

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

Модель DWH описана на **99,9%** и 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

qwen3-235b

Вот некоторые таблицы, которые могут пригодиться:

▼ prod_v_dwh_rep.query_launch 



Детализированные данные о запусках запросов GP

Columns:

log_flg smallint

query_launch_dt date

engine_code character varying(8)

node_launch_rk character varying(35)

start_dttm timestamp without time zone

process_launch_rk character varying(64)

query_launch_id numeric(25,0)

query_txt text

query_launch_rk uuid

query_type_nm text

[Show more](#)

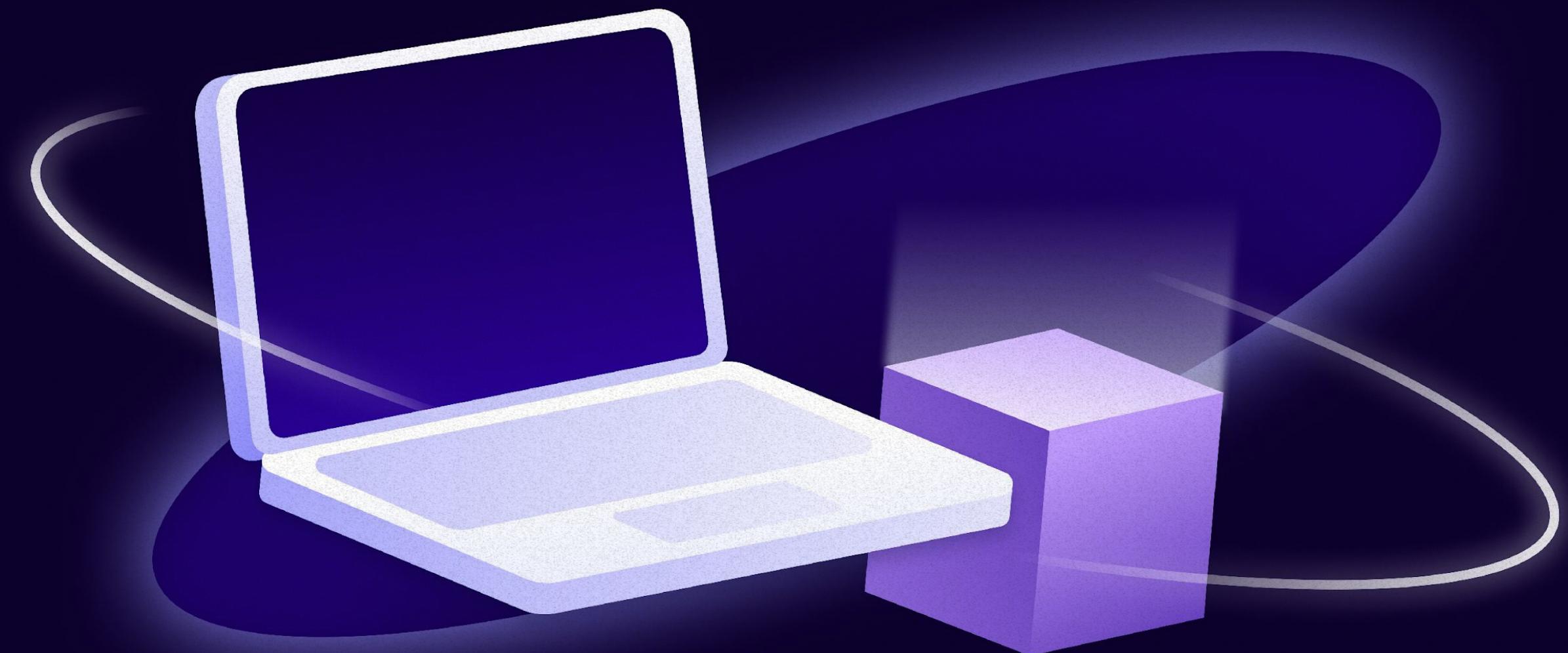
▶ prod_v_usermart.rtb_questionary 



AI Table Search
Семантический поиск
по таблицам Data
Platform

Исследование, ставшее продуктом

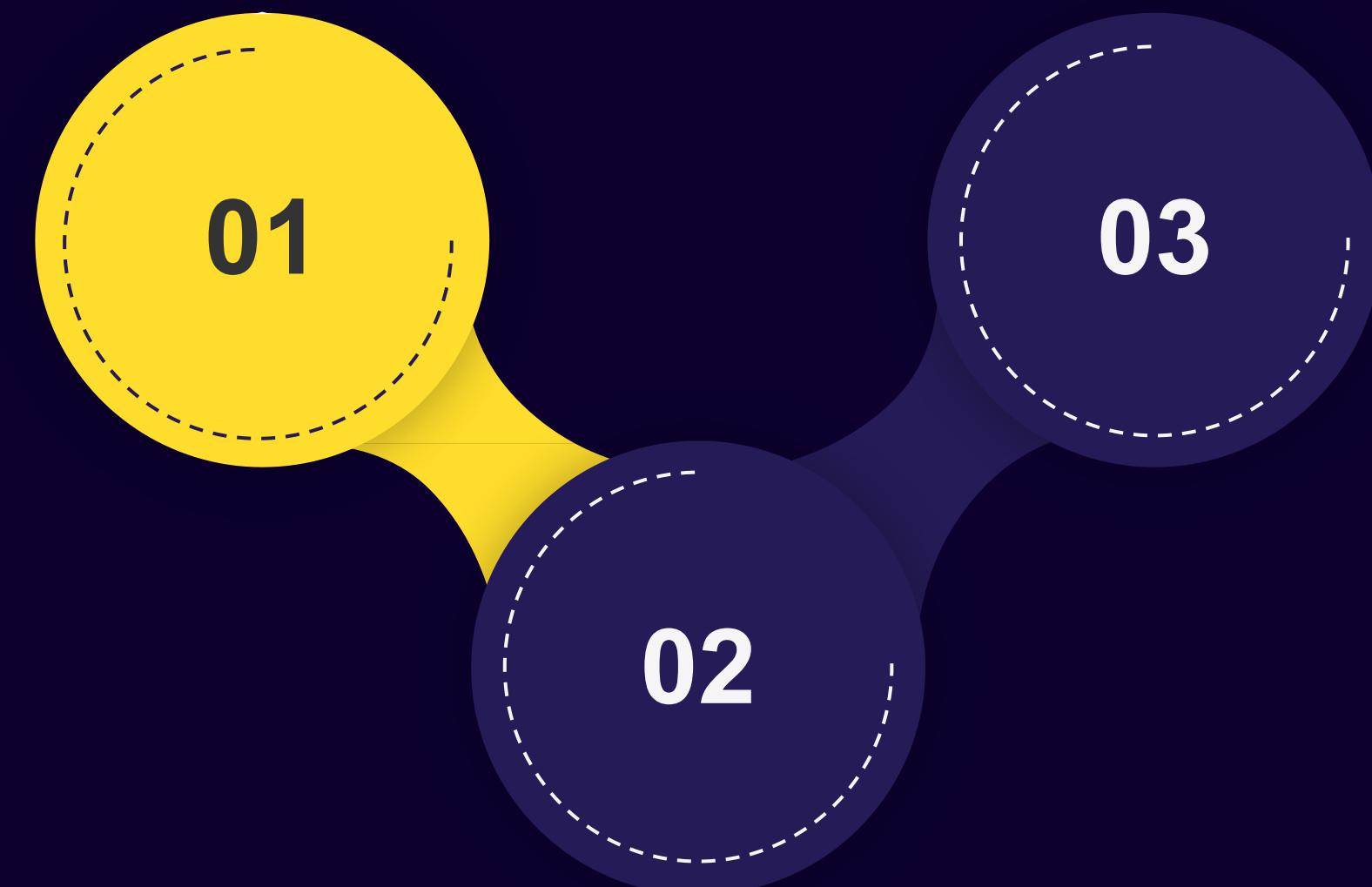
Или как RnD делает продукты из гипотез



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



Пошли в прод
Реализовав все
фич-реквесты
нет смысла
медлить...

Инструмент для генерации **autodesc** описаний таблиц

Исследование, ставшее
продуктом

usr_bi_core.autodesc_info Таблица

dap_limgpt not_reviewed userlabs_table

Структура Физические таблицы Связи Lineage & Usage

⚠ Если описание удовлетворяет требованиям, убери тег "not_reviewed"
Источник описания [Ulabs Autodesc](#)

✓ <https://helicopter.tcsbank.ru/notes/None>

Информация о генерации и оценке автоматических описаний объектов

Attributes

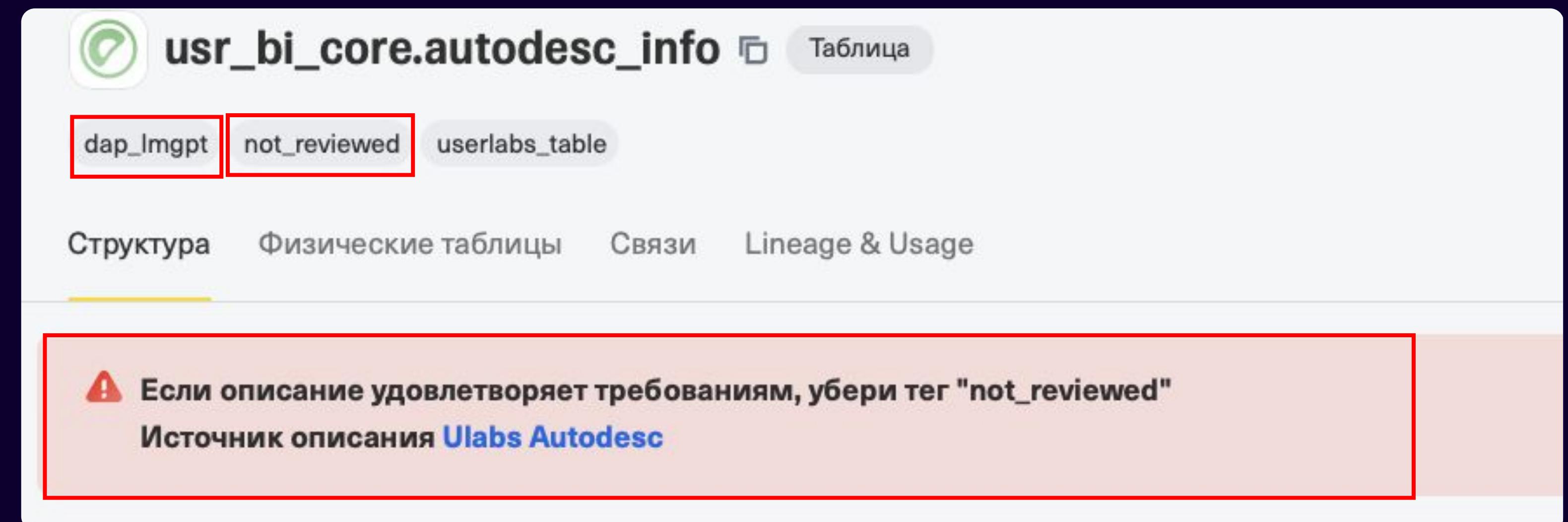
Владелец Oleg Sagitov
БД источника gp-idwh

Список колонок 24

Ключ	Тип ЧД	Имя	Описание
		pageid	Идентификатор страницы в системе
		contentid	Идентификатор содержимого страницы
		page_title	Заголовок страницы с описанием
		version	Версия содержимого страницы

Инструмент для генерации **autodesk** описаний таблиц

Исследование, ставшее
продуктом



usr_bi_core.autodesk_info Таблица

dap_Imgpt not_reviewed userlabs_table

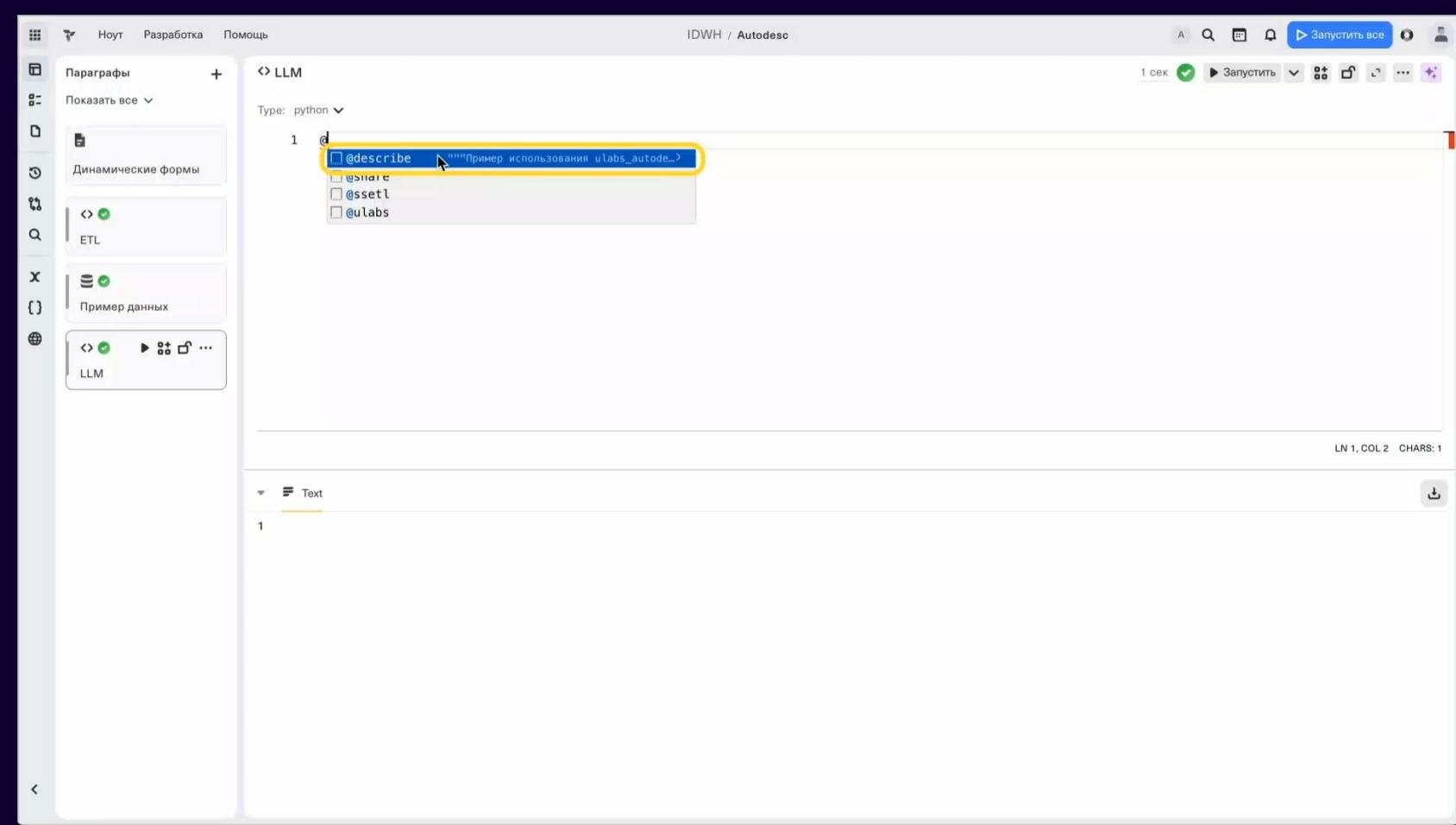
Структура Физические таблицы Связи Lineage & Usage

⚠ Если описание удовлетворяет требованиям, убери тег "not_reviewed"

Источник описания [Ulabs Autodesc](#)

autodesc

Исследование, ставшее продуктом

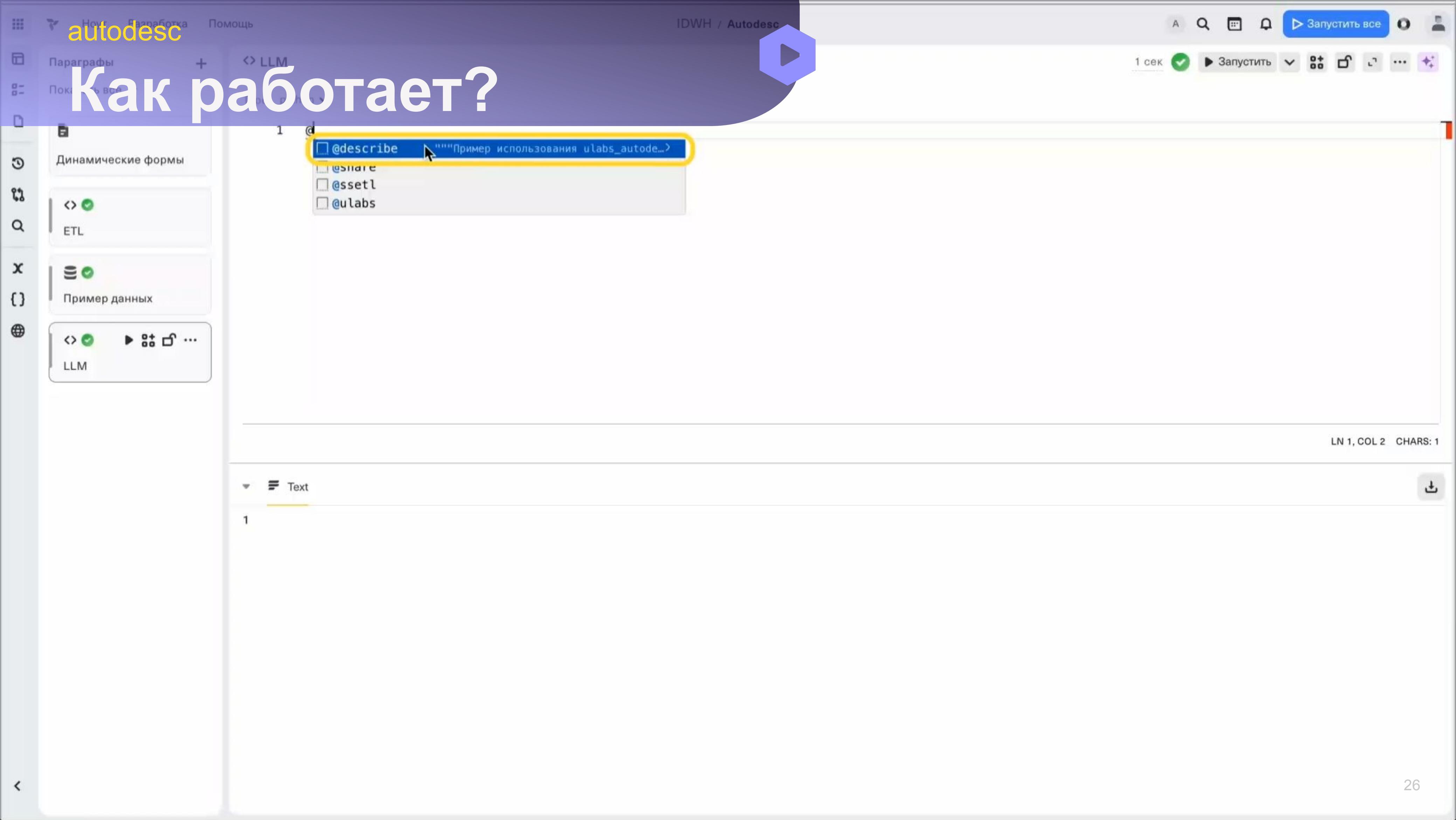


```
1 # @describe """Пример использования ulabs_autodesc"""
2
3 #ignore
4 #assetl
5 #ulabs
```



autodesk

Как работает?



1 @

1 @ @describe """Пример использования ulabs_autode...

1 @ @share

1 @ @assetl

1 @ @ulabs

Динамические формы

ETL

Пример данных

LLM

LN 1, COL 2 CHARS: 1

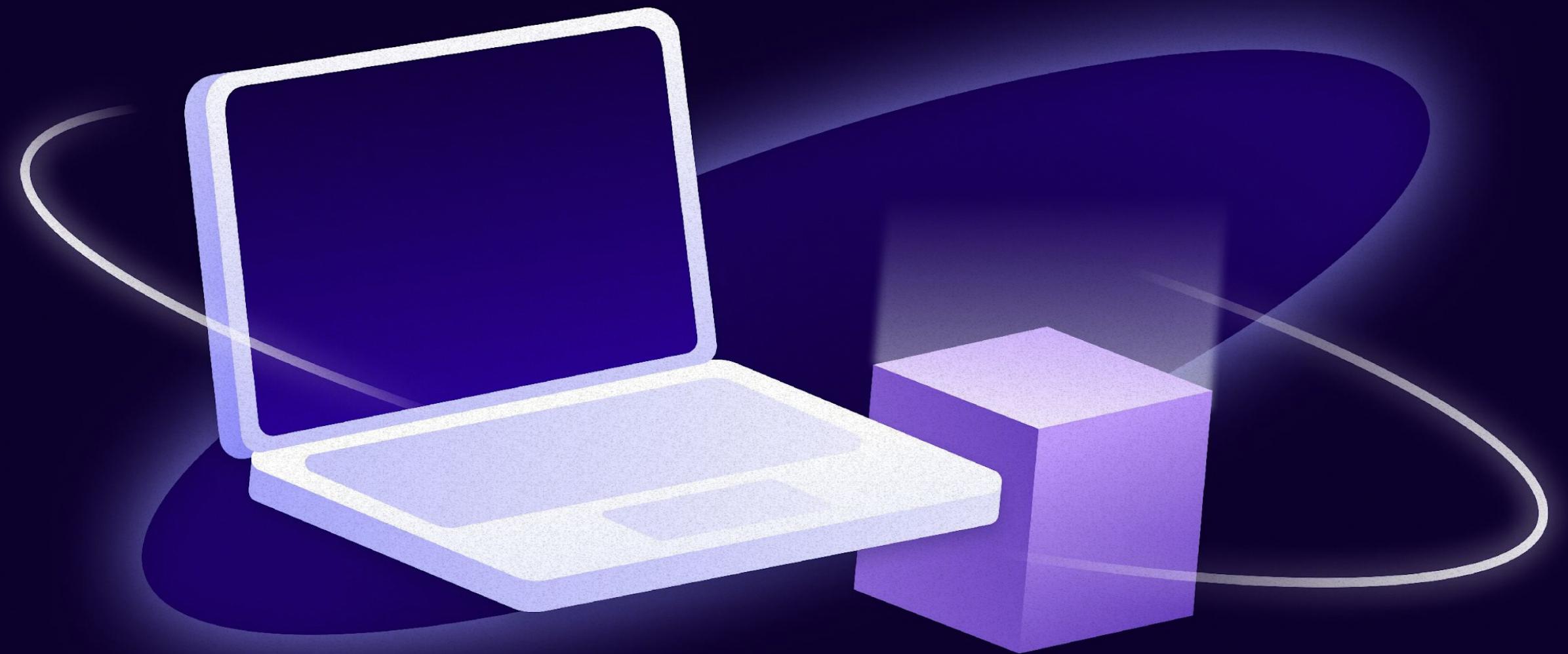
Text

1

26

Техническая реализация

autodesk



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:

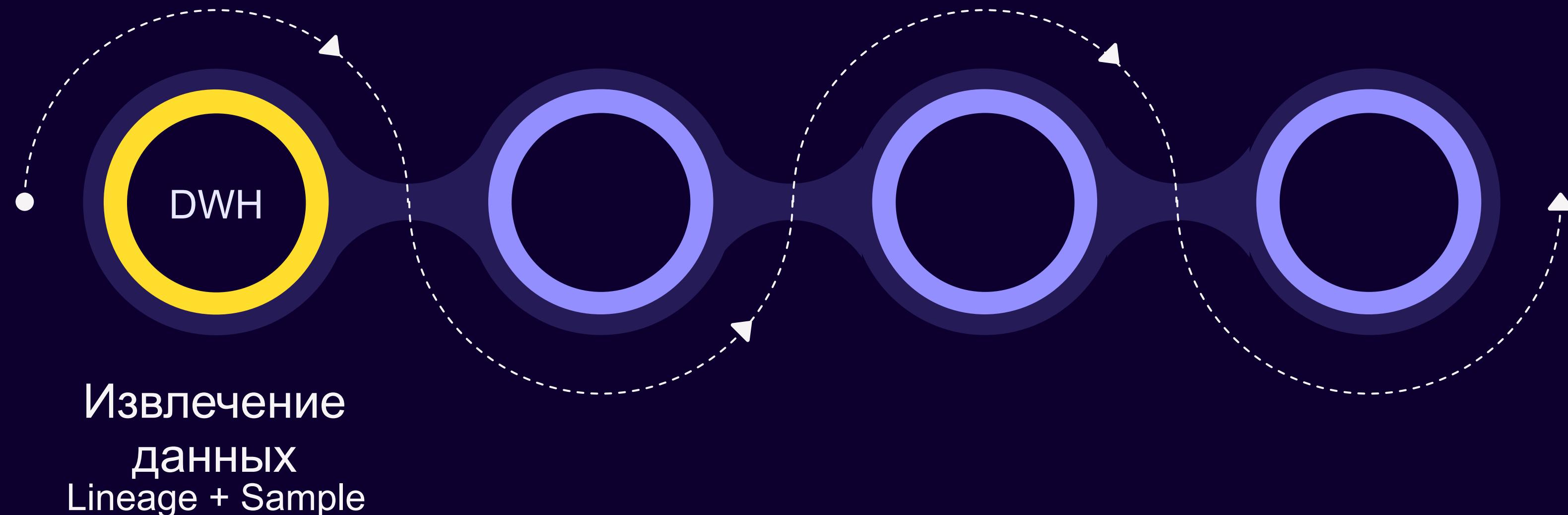
- {"term": "**Клиент**", "content": "Клиент банка или других продуктов холдинга \"Т\" (инвестиции, страхование, мобайл).", "synonyms": "Пользователь приложения, Участник, Покупатель"}

autodesk

Механизм описани я



Механизм описания



Механизм описания

```
{  
  notebook_url: "https://helicopter.tcsbank.ru/notes/1181370"  
  columns_desc: [...] 100 Items  
  tables_desc: [  
    0: {...} 3 Items  
    1: {  
      src_entity_name: "prod_v_sse.wiki_ulabs"  
      src_entity_desc: "История версий страниц вики с метками и флагами актуальности"  
      notebook_url: "https://helicopter.tcsbank.ru/notes/1181370"  
    }  
    2: {...} 3 Items  
    3: {...} 3 Items  
    4: {...} 3 Items  
    5: {  
      src_entity_name: "prod_v_chrono_llm_langfuse.traces_public"  
      src_entity_desc: "Информация о таблице ассистентов"  
      notebook_url: "https://helicopter.tcsbank.ru/notes/1181370"  
    }  
  ]  
}
```

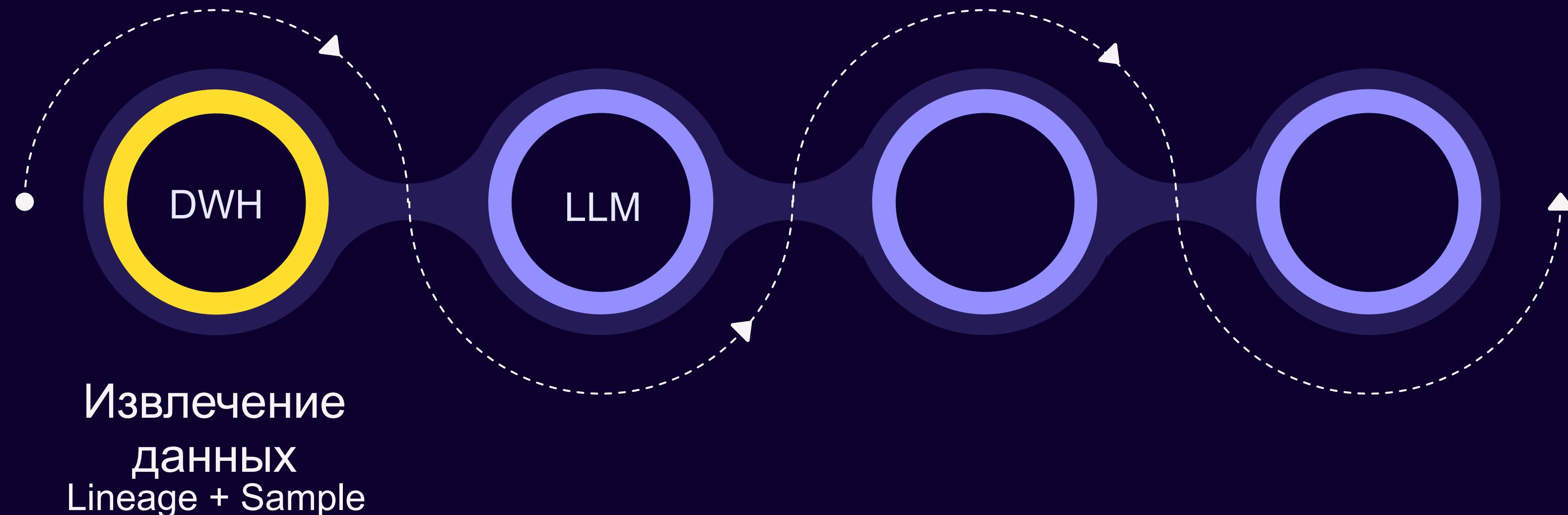
Metadata

```
{  
}
```

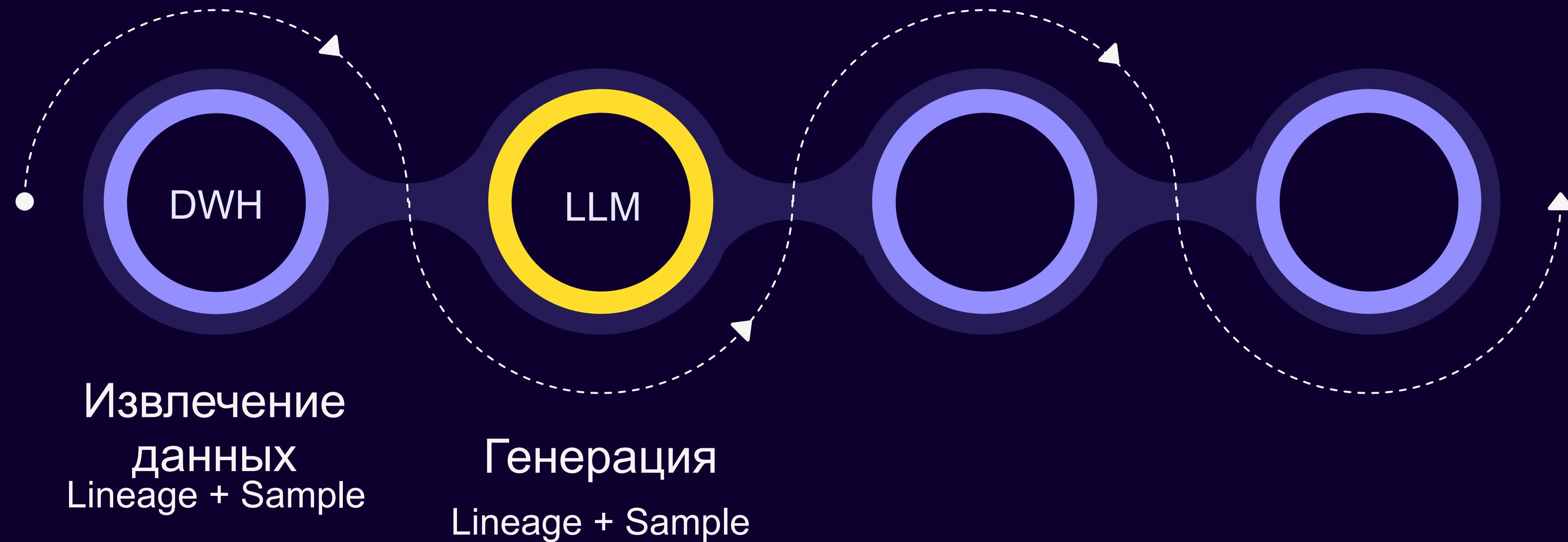
Lineage + Sample

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

Механизм описания



Механизм описания



```
{~  
  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"  
    }
```

autodesc

Генерация

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

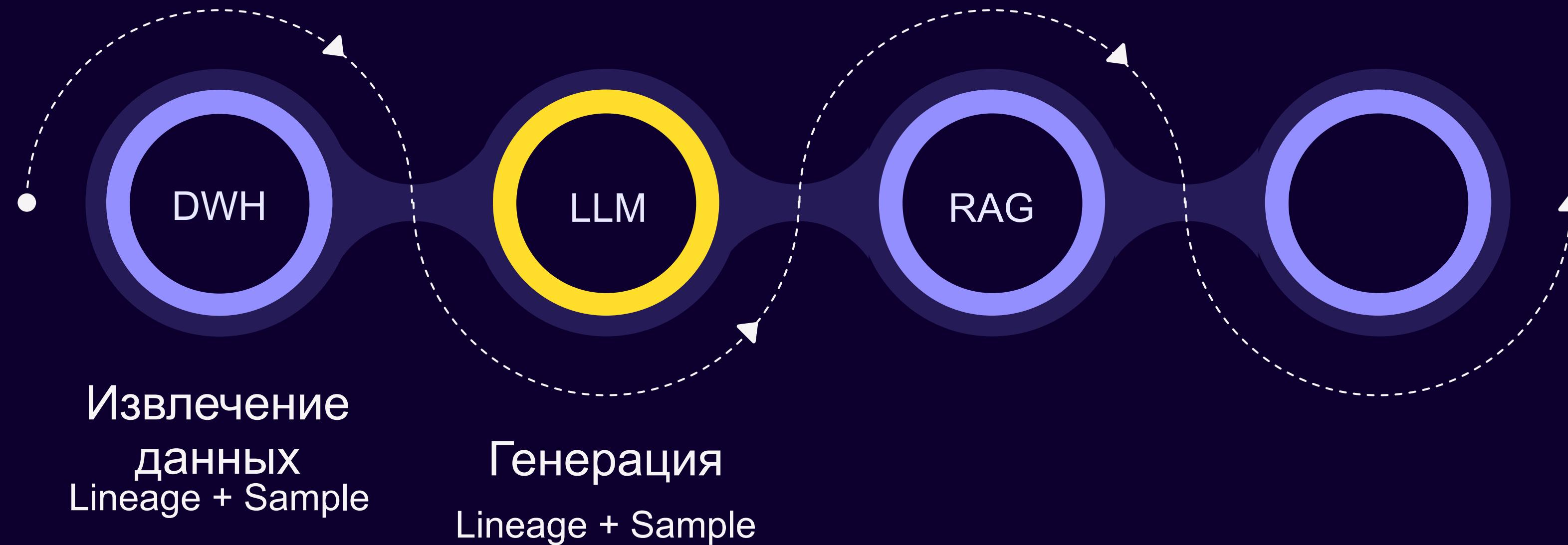


Сэмпл данных позволяет не путаться при ломанном именовании столцов



Lineage позволяет точнее выявлять взаимосвязь между данными

Механизм описания



Механизм описания



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.



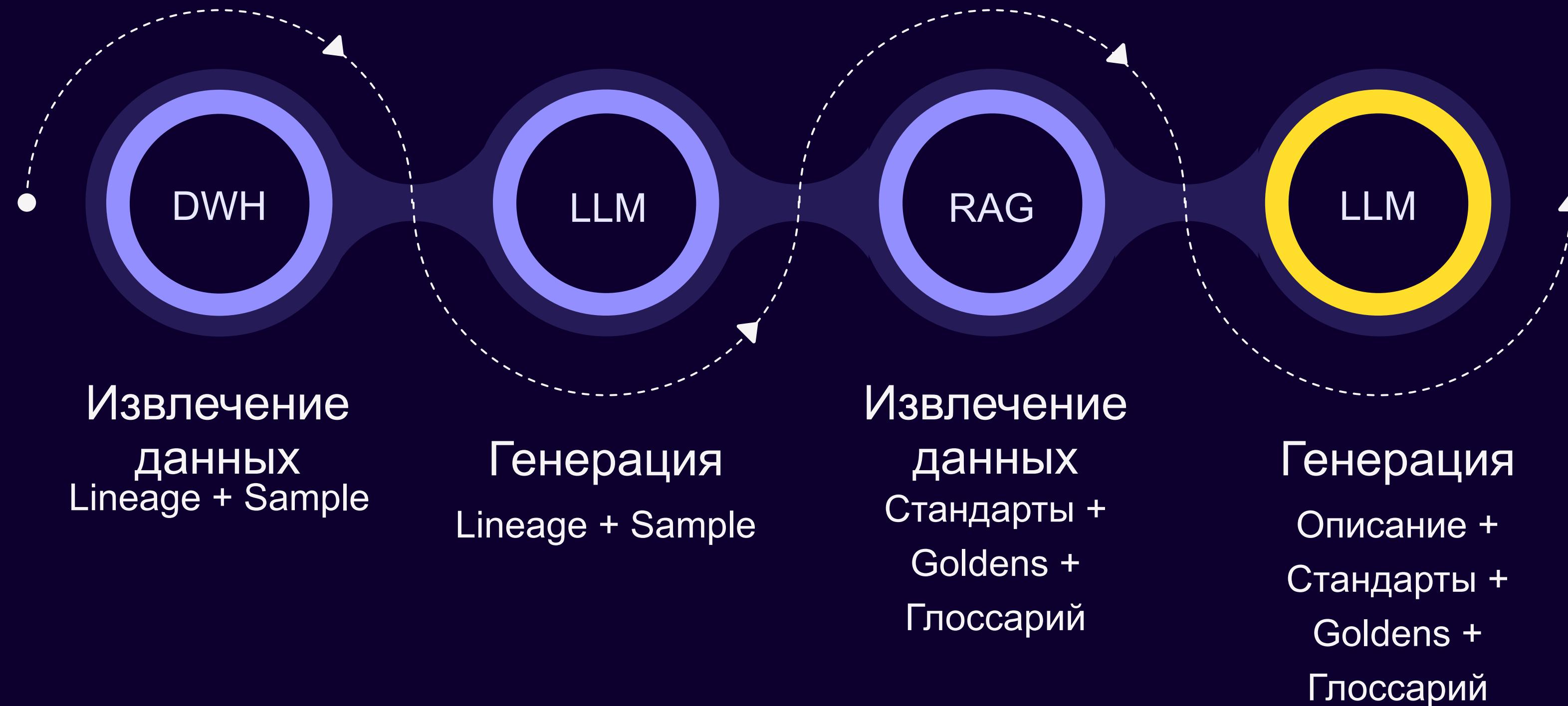
Валидация бизнеса

Бизнес явно выделил ценность объекта и связал его с терминами

Механизм описания



Механизм описания



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 ОЕІ 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



```
Preview Formatted JSON
Assistant

{
  table_desc: "Информация о версиях страниц с автоматическим описанием и метками"
  columns: [
    0: {
      column_nm: "pageid"
      column_desc: "Идентификатор страницы в системе"
    }
    1: {
      column_nm: "contentid"
      column_desc: "Идентификатор содержимого страницы"
    }
    2: {
      column_nm: "page_title"
      column_desc: "Заголовок 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: "Время начала автогенерации"
    }
    10: {...} 2 Items
  ]
}
```

- Уточнение pageid как идентификатора в системе

- Конкретизация contentid до «страницы»

- Сокращение и применение терминологии **autodesc_start_dttm**

Тест на проде

review

ssistant

```
{  
  table_desc: "Информация о версиях страниц с автоматиче-  
  columns: [  
    0: {  
      column_nm: "pageid"  
      column_desc: "Идентификатор страницы в системе"  
    }  
    1: {  
      column_nm: "contentid"  
      column_desc: "Идентификатор содержимого страницы"  
    }  
    2: {  
      column_nm: "page_title"  
      column_desc: "Заголовок страницы"  
    }  
    3: {  
      column_nm: "version"  
      column_desc: "Версия страницы"  
    }  
  ]  
}
```

- Уточнение **pageid** как идентификатора в системе

- Конкретизация **contentid** до «страницы»

- Сокращение и применение терминологии **autodesc_start_dttm**

Тест на проде



```
Preview Formatted JSON
Assistant

{
  table_desc: "Информация о версиях страниц с автоматическим описанием и метками"
  columns: [
    0: {
      column_nm: "pageid"
      column_desc: "Идентификатор страницы в системе"
    }
    1: {
      column_nm: "contentid"
      column_desc: "Идентификатор содержимого страницы"
    }
    2: {
      column_nm: "page_title"
      column_desc: "Заголовок 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: "Время начала автогенерации"
    }
    10: {...} 2 Items
  ]
}
```

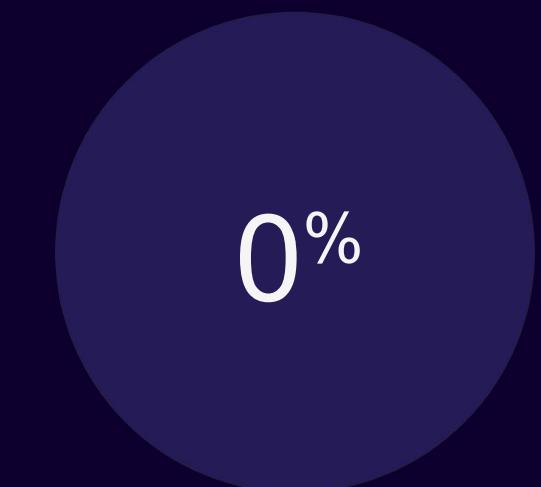
- Уточнение pageid как идентификатора в системе

- Конкретизация contentid до «страницы»

- Сокращение и применение терминологии **autodesc_start_dttm**

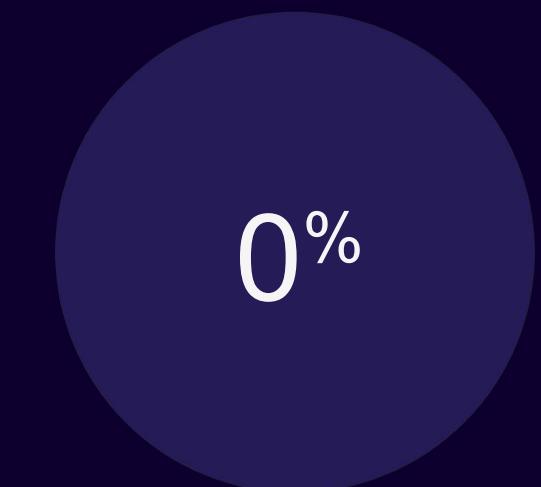
Online Eval

Тест на проде



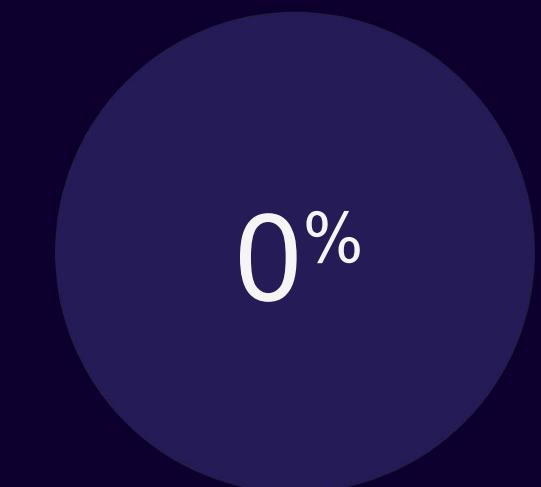
Cosine Similarity

Медиана матрицы косинусного сходства для N генераций.



ContextPrecision@k

Классика из фреймворка RAGAS.
Метрикой отвечаем на вопрос:
«Нашлось ли что-то полезное?»



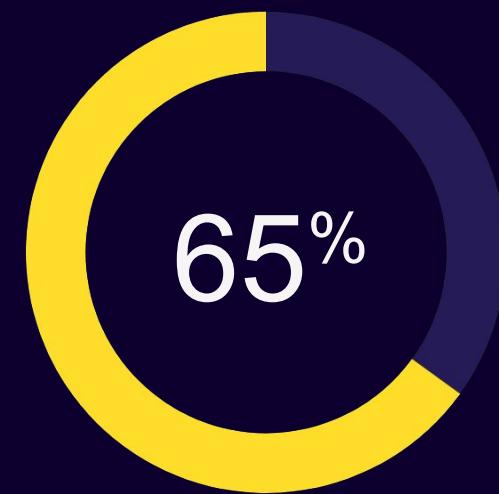
Answer Faithfulness

Немного трансформированная RAGAS-метрика.

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

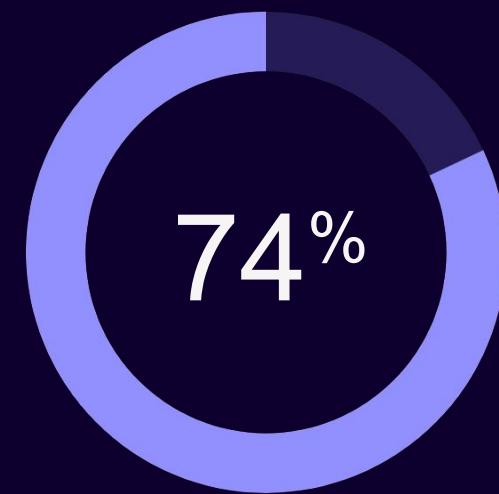
Online Eval

Тест на проде



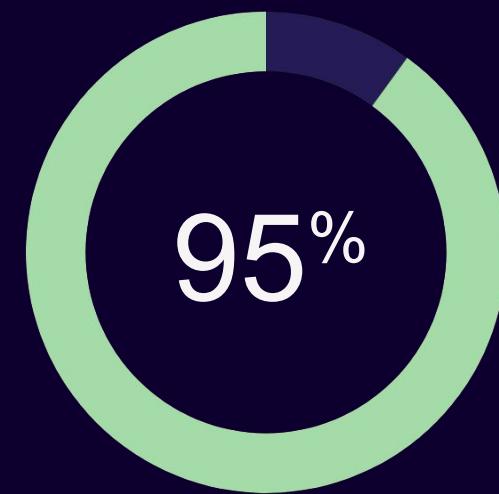
Cosine Similarity

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.

Moving average over time



Наблюдаемость

Позволяет без лишних приседаний
увидеть резкое падение метрик день ото
дня

```
# score_precision (api)
```

0.0000 ⓘ

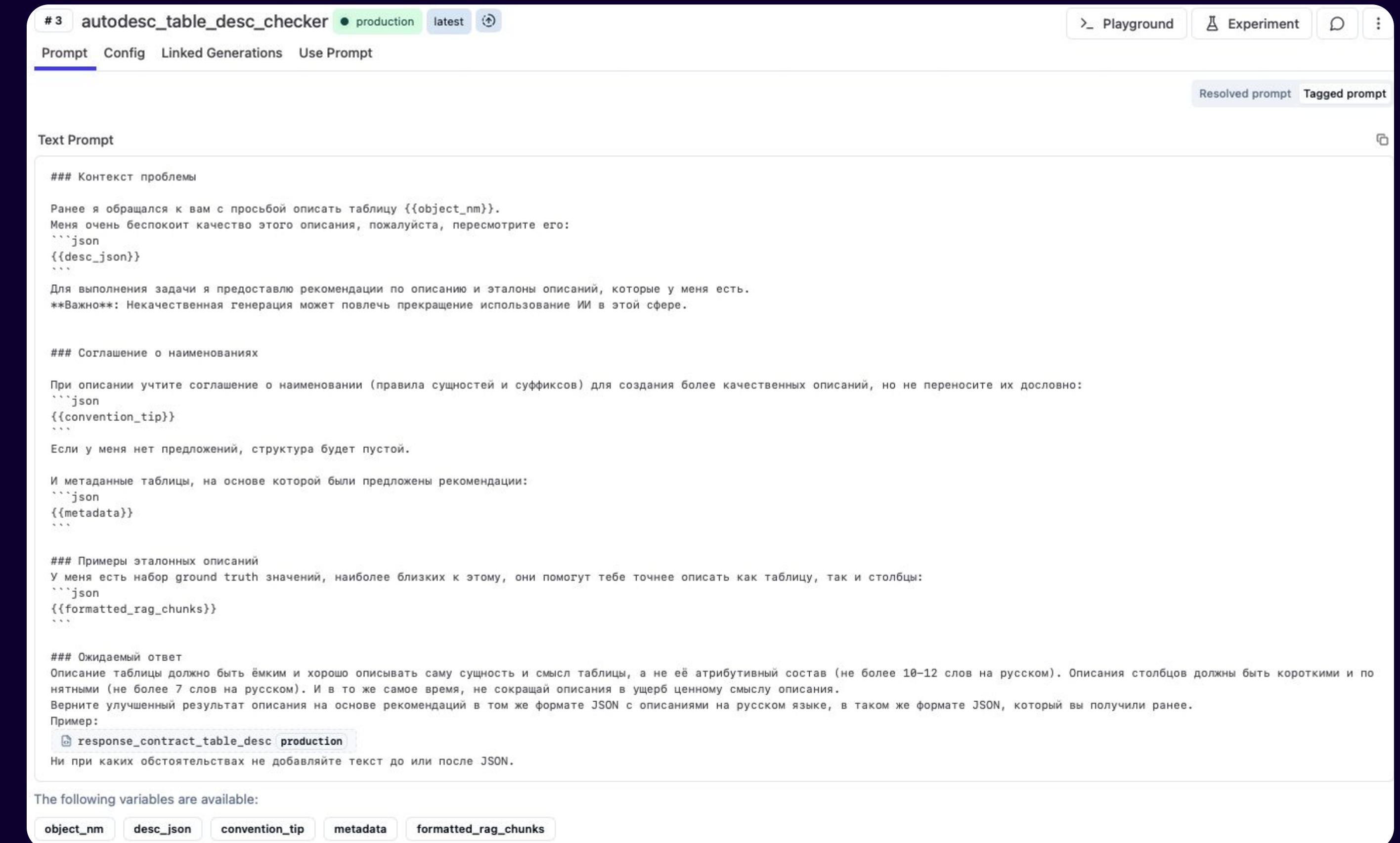
Контексты: 4, Полезные: 0

Прозрачность

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

Управляемость

Prompt-Management



The screenshot shows the Prompt-Management interface for the `autodesk_table_desc_checker` model in production. The interface includes a navigation bar with tabs for `Prompt`, `Config`, `Linked Generations`, and `Use Prompt`. The `Prompt` tab is selected. On the right, there are buttons for `Resolved prompt` and `Tagged prompt`. The main content area is titled `Text Prompt` and contains the following text:

```
### Контекст проблемы
Ранее я обращался к вам с просьбой описать таблицу {{object_nm}}.
Меня очень беспокоит качество этого описания, пожалуйста, пересмотрите его:
```json
{{desc_json}}
```
Для выполнения задачи я предоставлю рекомендации по описанию и эталоны описаний, которые у меня есть.
**Важно:** Некачественная генерация может повлечь прекращение использования ИИ в этой сфере.

### Соглашение о наименованиях
При описании учтите соглашение о наименовании (правила сущностей и суффиксов) для создания более качественных описаний, но не переносите их дословно:
```json
{{convention_tip}}
```
Если у меня нет предложений, структура будет пустой.

И метаданные таблицы, на основе которой были предложены рекомендации:
```json
{{metadata}}
```

### Примеры эталонных описаний
У меня есть набор ground truth значений, наиболее близких к этому, они помогут тебе точнее описать как таблицу, так и столбцы:
```json
{{formatted_rag_chunks}}
```

### Ожидаемый ответ
Описание таблицы должно быть ёмким и хорошо описывать саму сущность и смысл таблицы, а не её атрибутивный состав (не более 10-12 слов на русском). Описания столбцов должны быть короткими и понятными (не более 7 слов на русском). И в то же самое время, не сокращай описания в ущерб ценному смыслу описания.
Верните улучшенный результат описания на основе рекомендаций в том же формате JSON с описаниями на русском языке, в таком же формате JSON, который вы получили ранее.
Пример:
response_contract_table_desc production
```

Ни при каких обстоятельствах не добавляйте текст до или после JSON.

The following variables are available:

`object_nm` `desc_json` `convention_tip` `metadata` `formatted_rag_chunks`

Управляемость

Модульность промптов

Ожидаемый ответ

Описание таблицы должно быть ёмким и хорошо описывать саму сущность и смысл таблицы, а не её атрибутивный составными (не более 7 слов на русском). И в то же самое время, не сокращай описания в ущерб ценному смыслу описания.

Верните улучшенный результат описания на основе рекомендаций в том же формате JSON с описаниями на русском языке.

Пример:

response_contract_table_desc production

НИ при каких обстоятельствах не добавляйте текст до или после JSON.

Eval + Prompt-Management

Prompt `autodesc_table_desc_checker` ⓘ

Versions Metrics

Columns 4/15

| Version | Labels | Trace: # answer_faithfulness (api) | Trace: # cosine_similary (api) |
|---------|------------|------------------------------------|--------------------------------|
| 3 | latest | | |
| 2 | production | Ø 0.9565 | Ø 0.9900 |
| 1 | | | |

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.

Влияние

CSAT

4.4

На основе опроса
более чем 150
респондентов.

WAU

59

%

От всех аналитиков
создающих описания
в DWH (210+)

Влияние

CSAT

4.4

На основе опроса
более чем 150
респондентов.

WAU

59

%

От всех аналитиков
создающих описания
в DWH (210+)

Доля AI-контента

42

%

Уже почти половина
пользовательских-
таблиц описаны с
помощью AI.

Влияние на пользователе й

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 пользователе й

Как повлияло на взаимодействие с платформой?



Применение для дашбордов и
песочниц

- Автоматизация покрытия пользовательских доменов описаниями
- Облегчение и развитие поиска по Встраивание в лайпайн ETL-аналитическим артефактам разработки:
- Ускорение автоматизированного ревью
- Упрощение актуализации при доработках

Влияние на пользователя

Как повлияло на взаимодействие с платформой?



Применение для дашбордов и песочниц

- Автоматизация покрытия пользовательских доменов описаниями
- Облегчение и развитие поиска по Встраивание в лайпайн ETL-аналитическим артефактам разработки:
- Ускорение автоматизированного ревью
- Упрощение актуализации при индикации фич для каталогизации:
- Оптимизация взаимодействия с платформой
- Выделение аналитической ценности



Спасибо!

