

# **Итоговый проект на тему: «Мультимодальная RAG-система по отчету Сбера»**

Акимов Дмитрий Андреевич

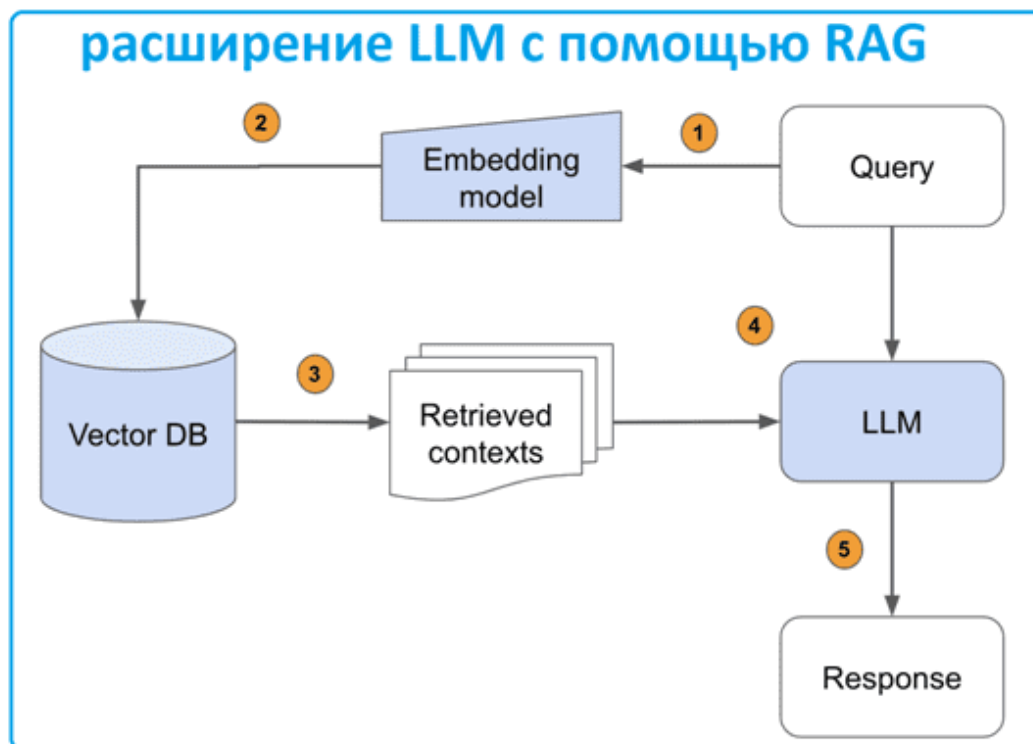
Шумов Александр Владимирович

# Актуальность темы и ее проблематика

## работа базовой LLM



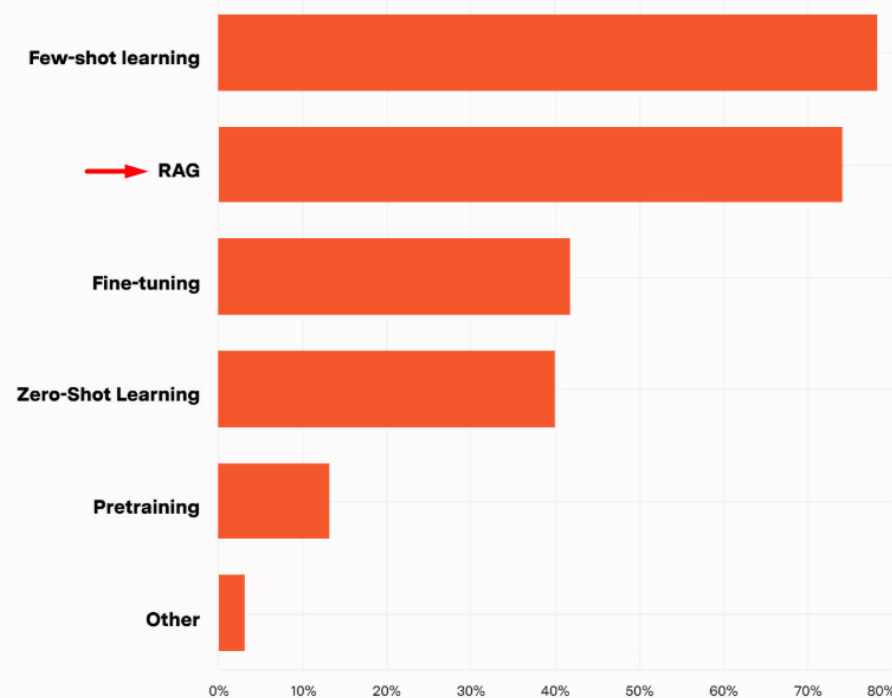
## расширение LLM с помощью RAG



Engineering and Infrastructure

## The 2025 AI Engineering Report

What techniques are you using to customize your AI systems?



# State-of-the-art RAG техники

habr.com/ru/articles/893356/



Хабр

КАК СТАТЬ АВТОРОМ

Разработчикам плюс вайб

Моя лента Все потоки Разработка Администрирование Дизайн Менеджмент Маркетинг Научпоп

IlyaRice 22 мар в 11:54

Как я победил в RAG Challenge: от нуля до SoTA за один конкурс

Средний 23 мин 35K

Искусственный интеллект, Natural Language Processing\*, Data Engineering\*, Машинное обучение\*

Конкурс на деньги (как и в случае с Kaggle) – часто лучший способ определить state-of-the-art техники.

В чём суть RAG Challenge?

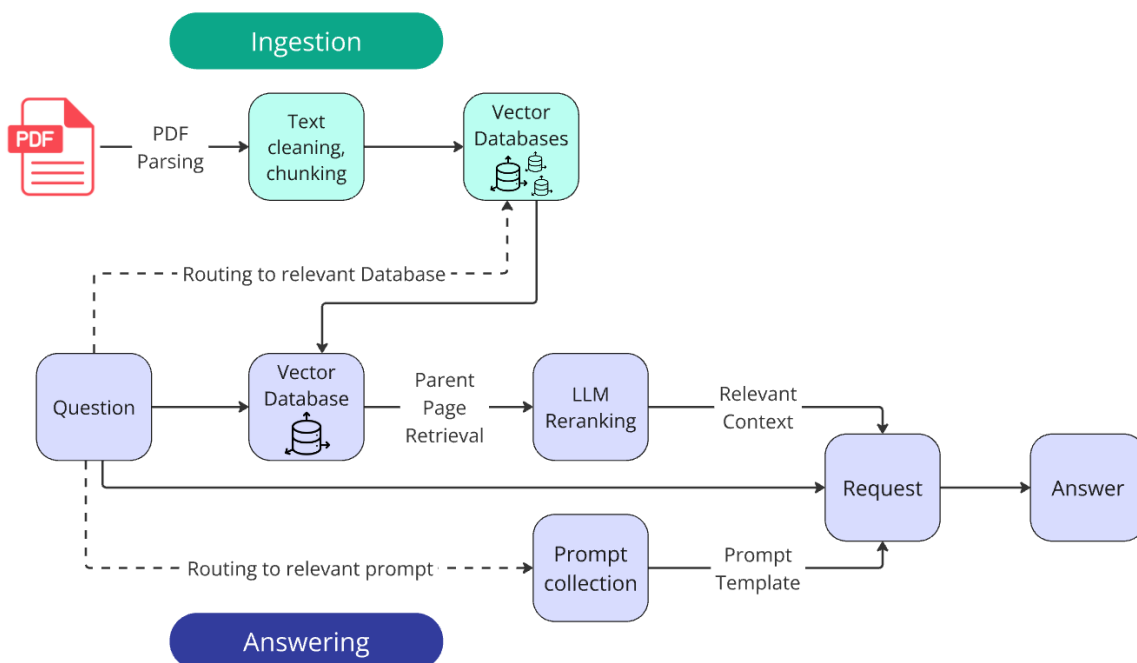
Нужно создать вопросно-ответную систему на основе годовых отчётов компании. Если коротко, то в день конкурса:

1. Выдаётся 100 годовых отчётов по случайно выбранным компаниям и 2.5 часа на их парсинг и составление базы данных. Отчёты представляют из себя PDF размером до 1000 страниц.
2. После этого генерируется 100 случайных вопросов (по заранее известным шаблонам), на которые система должна как можно быстрее ответить.

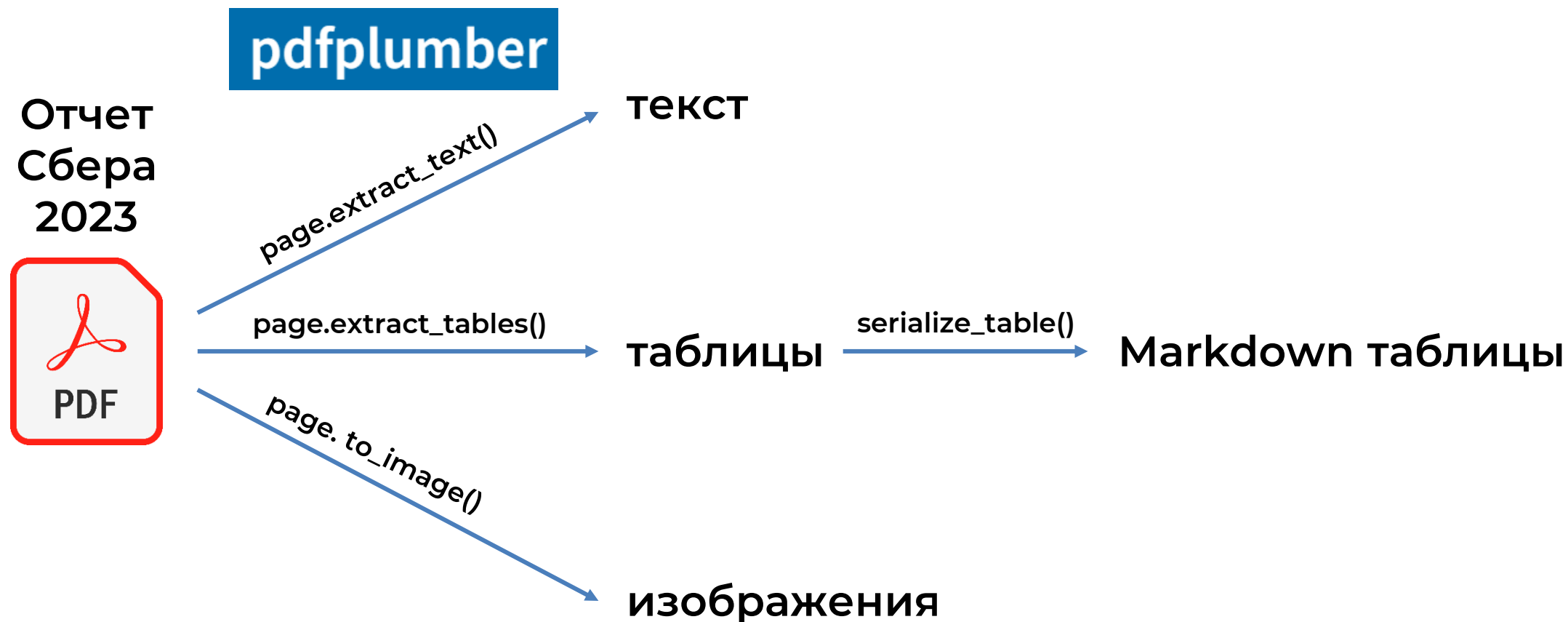
Все вопросы должны иметь однозначный ответ:

- Да/Нет;
- название компании (или нескольких компаний);
- названия управляющих позиций, выпущенных продуктов;
- размер той или иной метрики: выручка, количество магазинов и т.д.

Каждый ответ должен сопровождаться ссылками на страницы с ответом в качестве доказательства, что система по честному нашла ответ и не сгаллюцинировала.

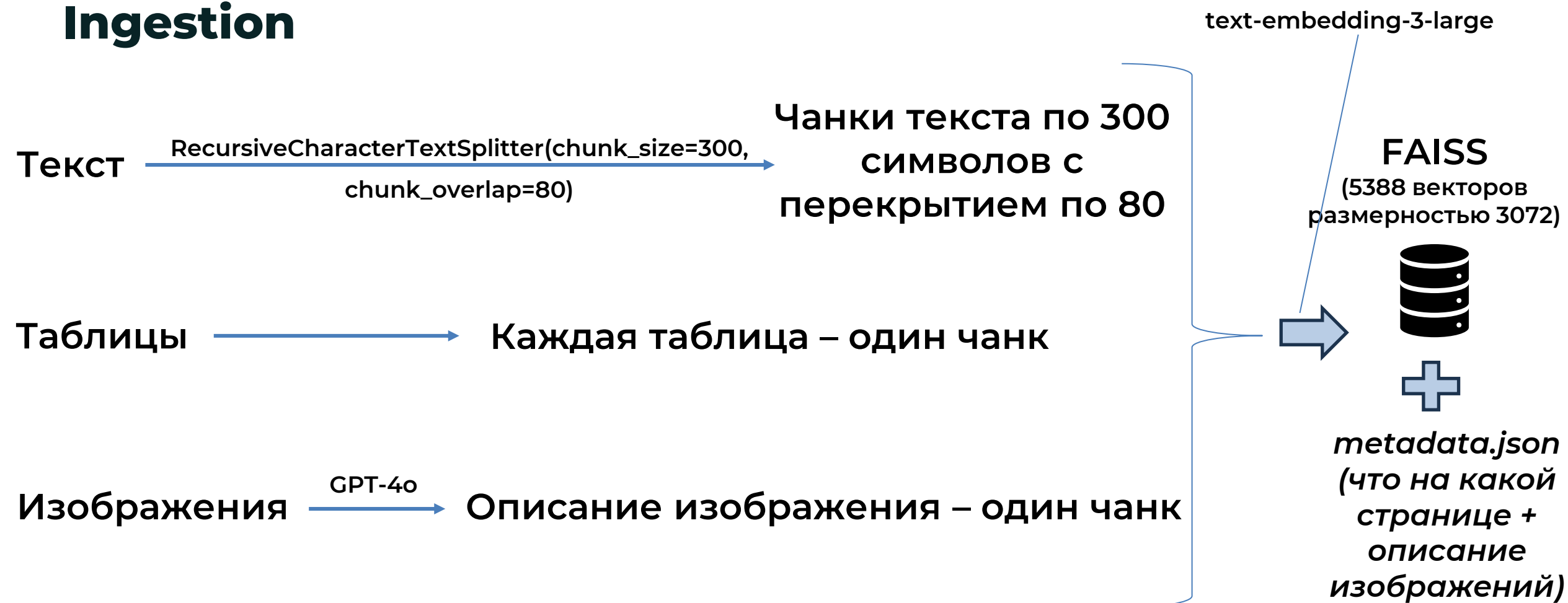


# Parsing



Победитель Enterprise RAG Challenge использовал Docling, арендовав сервер на Runprod – мы сделали так же, но долго было разбираться с конфликтом зависимостей, и вернулись на pdfplumber... А unstructured работал очень долго.

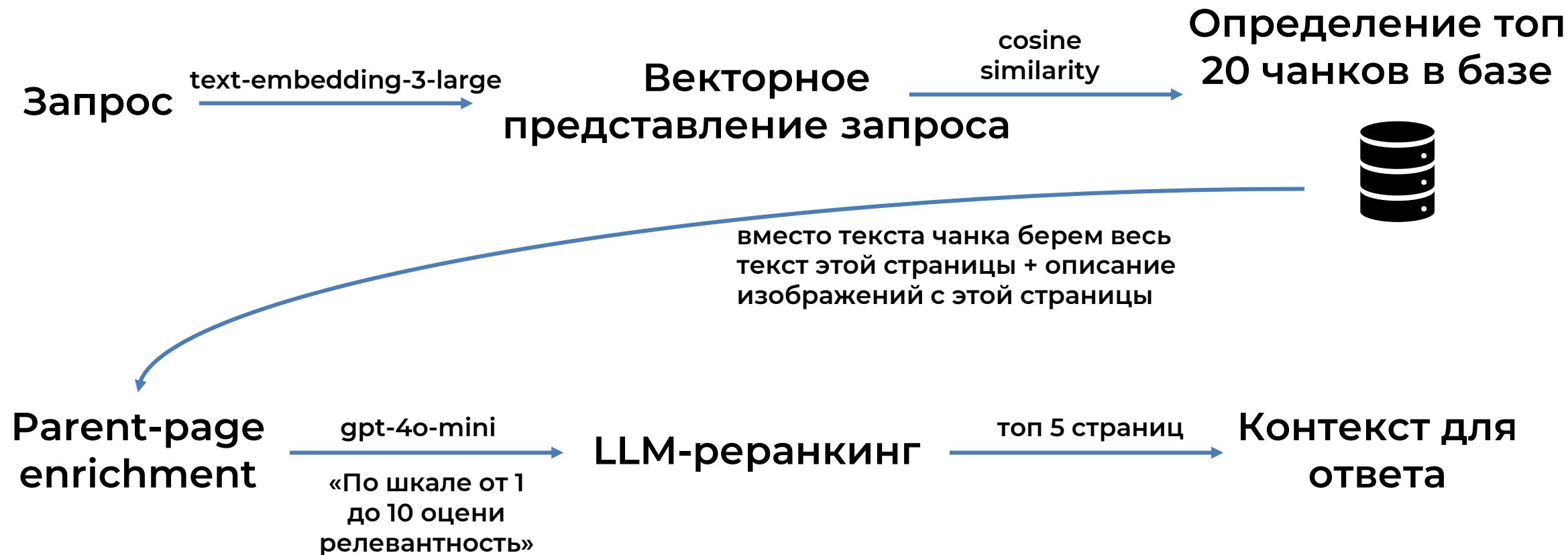
# Ingestion



Победитель Enterprise RAG Challenge использовал тот же пайплайн, но делил на чанки текст вперемешку с таблицами и использовал перекрытие 50.



# Retrieval



Победитель Enterprise RAG Challenge использовал тот же пайплайн, но брал топ 30 чанков / топ 10 страниц, поскольку у него было больше документов на вход.

# Augmentation + Generation

Ты эксперт по финансовым отчетам.  
Ответь, используя Chain-of-Thoughts.

ВОПРОС: (запрос пользователя)

КОНТЕКСТ: (ранее найденный контекст)

Ответь в требуемом формате (мышление,  
ответ, уверенность, источники) в json.

GPT-4o, pydantic

Ответ

если не в требуемом формате

Ты эксперт по финансовым отчетам.  
Ответь, используя Chain-of-Thoughts.

ВОПРОС: (запрос пользователя)

КОНТЕКСТ: (ранее найденный контекст)

Ответь в требуемом формате (мышление,  
ответ, уверенность, источники).

GPT-4o

Ответ

# Demo

Задайте вопрос

Ваш вопрос

Каковы основные финансовые показатели Сбера за 2023 год?

Поиск

Ответ системы

- Итого обязательств составило 43 723 млрд руб., увеличившись на 20,7% (страница 94).
- Итого собственных средств составило 6 584 млрд руб., увеличившись на 14% (страница 94).

4. **Качество активов:**

- Доля неработающих кредитов в кредитном портфеле снизилась до 2,2% (страница 95).

**РАССУЖДЕНИЯ:**

На основе предоставленных данных можно сделать выводы о значительном улучшении финансовых показателей Сбера в 2023 году по сравнению с предыдущим годом. Это выражается в росте рентабельности, увеличении чистой прибыли и операционных доходов, а также в снижении доли неработающих кредитов. Также наблюдается рост активов и обязательств, что свидетельствует о расширении деятельности банка.

**ВЫВОДЫ:**

Основные финансовые показатели Сбера за 2023 год демонстрируют значительное улучшение:

- Рентабельность капитала (ROE) составила 25,3%.
- Чистая прибыль увеличилась в 5,2 раза до 1 508,6 млрд руб.
- Операционные доходы выросли на 121,7% до 2 909,1 млрд руб.
- Активы увеличились на 25% до 52 307 млрд руб.
- Доля неработающих кредитов снизилась до 2,2%.

Источники и статистика

**Источники:**

**1.** Страница 95 (поиск: 0.725, релевантность: 10.0/10)

**Превью:** Финансовые показатели  
Показатели рентабельности  
%

2023

2022

Изменение

2021

Рентабельность среднегодовых активов  
(ROA)

3,2

0,7

2,5 п.п.

3,3

Рентабельность капитала (ROE)

25,3

5,2



## Выводы

- RAG – одно из основных применений LLM на сегодняшний день.
- В работе была реализована мультимодальная RAG-система, вдохновленная победным решением Enterprise RAG Challenge (2025 год).
- Основные используемые технологии: pdfplumber, RecursiveCharacterTextSplitter, FAISS, Parent-page enrichment, LLM reranking, Chain-of-Thoughts, pydantic, gradio, GPT-4o, text-embedding-3-large, gpt-4o-mini.
- Основные сложности: некачественное выделение и распознавание картинок (следует рассмотреть переход с pdfplumber на docling, а также уточнить промпт для GPT-4o и добавить автоматический повторный запрос в случае отказа создать описание картинки).
- Перспективы на будущее: целесообразно верифицировать качество собранного пайплайна путем построения автоматической оценки (RAGAS и др.).

## Список используемых источников/программных средств:

- ✔ Google Colab, HuggingFace
- ✔ «Как я победил в RAG Challenge: от нуля до SoTA за один конкурс» (<https://habr.com/ru/articles/893356/>)