

# **State-of-the art техники RAG на примере победного решения Enterprise RAG Challenge и его адаптации**

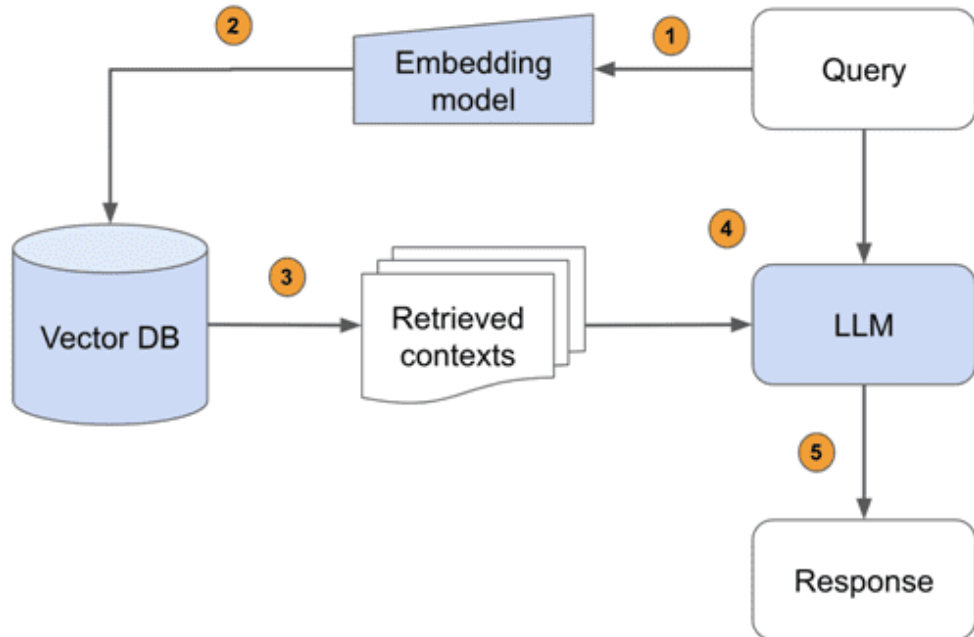
Акимов Дмитрий

# Актуальность RAG

## работа базовой LLM



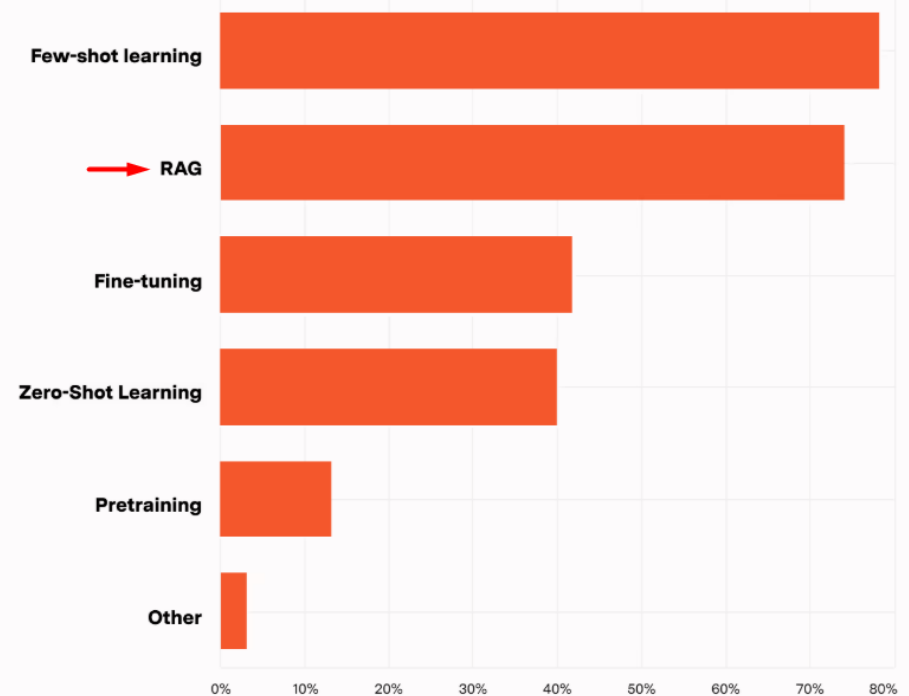
## расширение LLM с помощью RAG



Engineering and Infrastructure

## The 2025 AI Engineering Report

What techniques are you using to customize your AI systems?



# Retrieval Augmented Generation (на пальцах)

Вопрос пользователя

Сколько Васе лет?

$\{0; 0.6; 0\}$

нарезаем на части («чанки»)

переводим в векторы с  
помощью модели эмбедингов

## Наши данные

Вася играет в футбол.  
15-летний Вася не по годам умен. Вася хорош в математике.

Вася играет в футбол.

$\{0; 0; 0.8\}$

$|\{0*0; 0*0.6; 0.8*0\}|=0$

15-летний Вася не по годам умен.

$\{0; 0.7; 0\}$

$|\{0*0; 0.7*0.6; 0*0\}|=\mathbf{0.42}$

Вася хорош в математике.

$\{0.5; 0; 0\}$

$|\{0.5*0; 0*0.6; 0*0\}|=0$

ищем косинусное сходство

## Запрос LLM

Сколько Васе лет?  
Найденный контекст:  
15-летний Вася не по годам умен.

## Ответ LLM

Васе 15 лет



# Retrieval Augmented Generation (если погружаться)

HyDE

RAG-Fusion

Haystack

LlamaIndex

Langchain

Knowledge Graph

MMR TrueLens

RAGAS

Agentic RAG

Recall@k



BM25 + TF-IDF

Contextual  
Retrieval Neo4j

Cross-Encoder  
Reranking

HNSW  
IVF-Flat

AutoRAG

Multi-Query

ColBERT

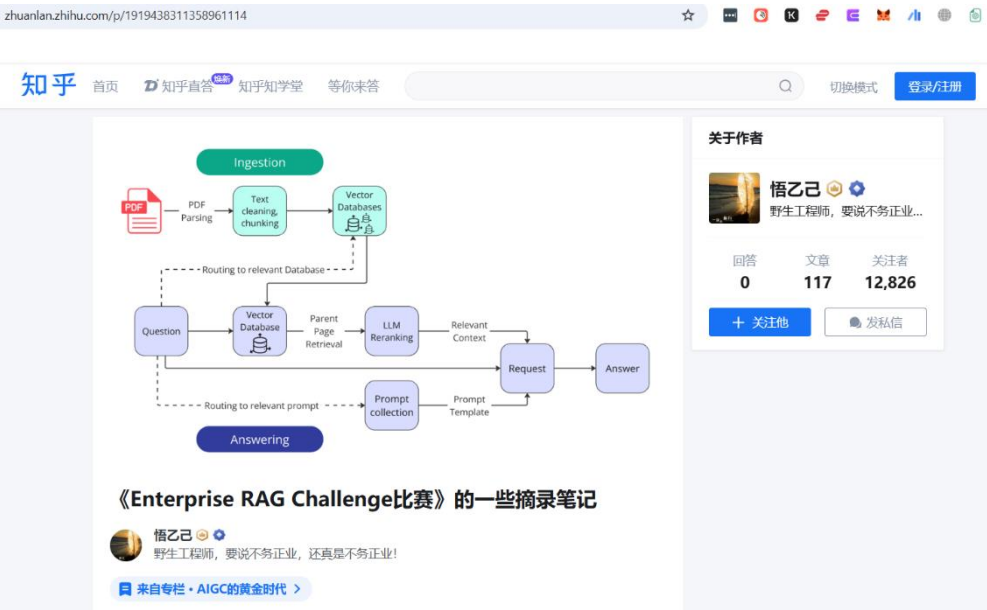
DeepEval

CRAG

# Enterprise RAG Challenge v2 (2025)



Статья получила международную популярность и была переведена на английский и китайский:



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.

# Enterprise RAG Challenge v2 (условия)

100 PDF годовых отчетов компаний (случайно отобранных)

100 случайных вопросов (Да/Нет; Числовые показатели и т.д.)

Оценивался Retrieval (верная ссылка в ответе) и Generation (верный ответ)

Итоговая метрика  $R/3 + G$  (т.е. верный ответ оценивался в 3 раза выше верного цитирования)



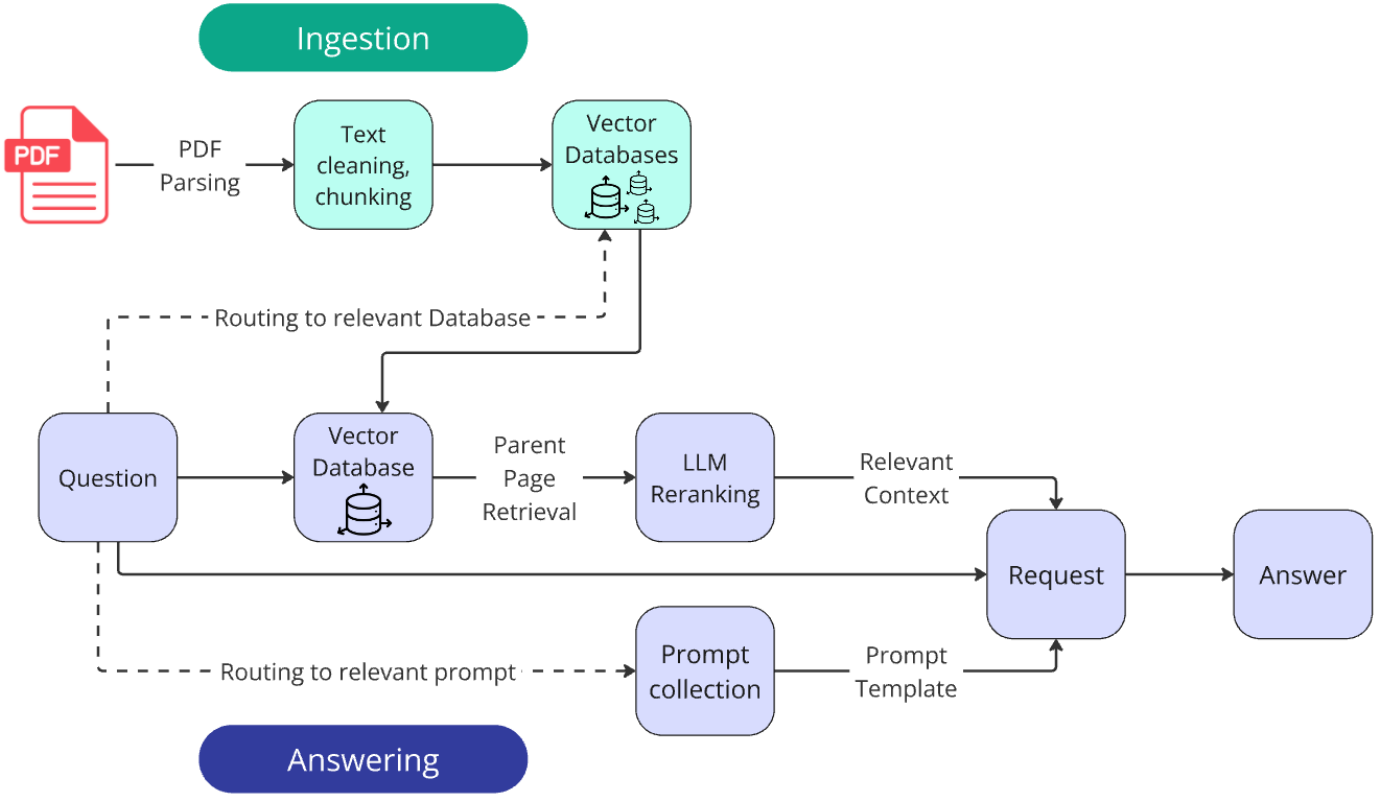


# Enterprise RAG Challenge v2. Победное решение (кратко)

Team / Experiment		Time	R/G	Score
1. ▶ Ilia Ris	🥇	49 min	83/81	123.7
2. ▶ Emil Shagiev	🥈	55 min	86/78	121.6
3. ▶ Dmitry Buykin	🥉	8 hours	81/76	117.5
4. ▶ Sergey Nikonov	🥈	30 hours	85/73	116.4
5. ▶ ScrapeNinja.net	🥈	23 hours	82/71	112.5
6. ▶ xsl777	🥈	16 hours	79/71	110.9
7. ▶ nikolay_sheyko(grably.tech)	🥈	25 hours	81/69	110.4
8. ▶ Felix-TAT	🥈	7 days	80/69	109.4
9. ▶ A.Rasskazov/V.Kalesnikau		30 hours	84/67	109.3
10. ▶ Dany the creator	🥈	3 hours	82/67	108.4
11. ▶ SergC	🥈	7 days	77/69	108.1
12. ▶ Swisscom Innovation Lab	🔒	21 hours	83/66	107.8
13. ▶ fomih	🥈	10 days	83/65	107.4
14. ▶ Al Bo		12 days	81/65	105.9
15. ▶ NumericalArt		8 days	70/70	105.3

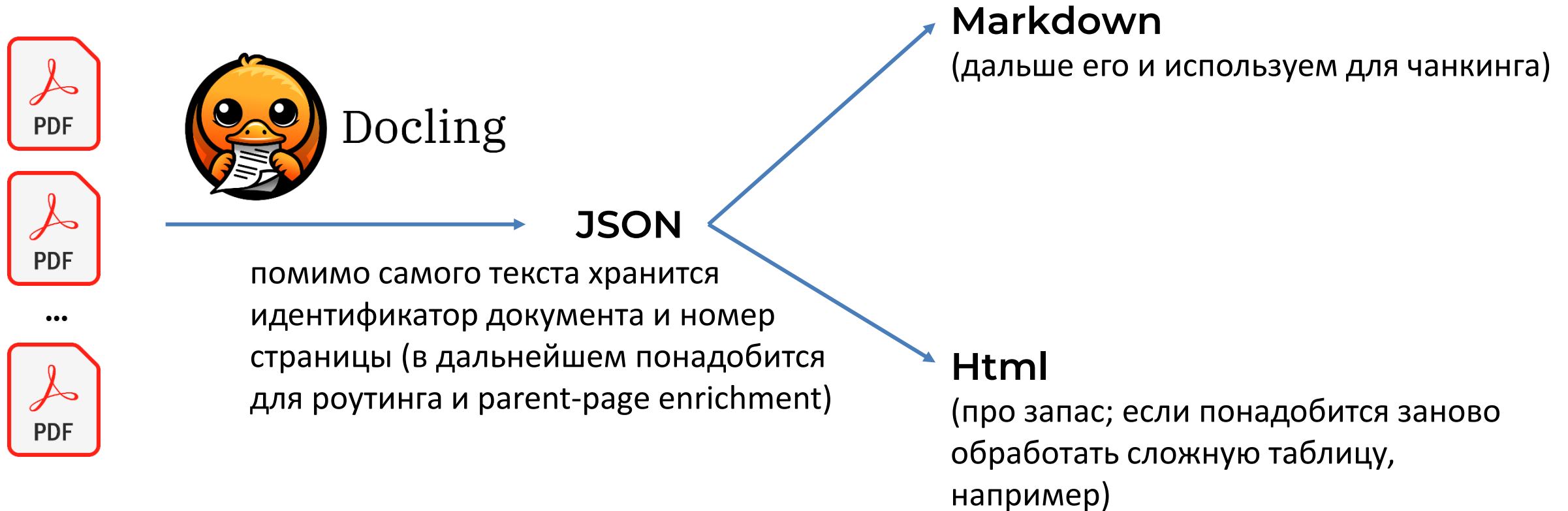
всего около 45 команд

## Схема победного решения Ilia Ris



На доп. треке IBM (использование open-source LLM Llama 2 70B) качество просело менее, чем на 2%!

# Parsing

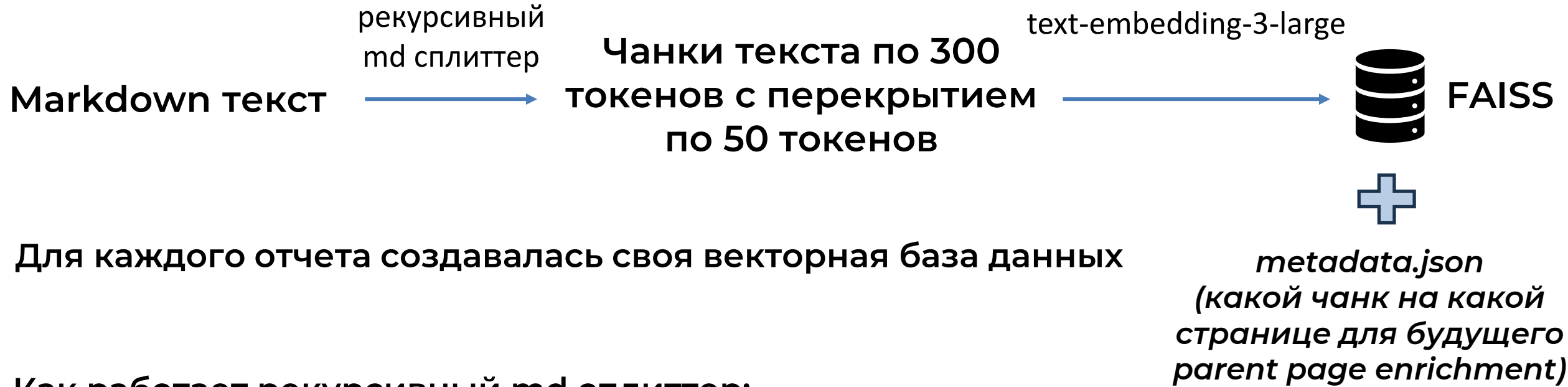


Илья использовал Docling, арендовав сервер на Runpod. Можно также использовать Unstructured, Marker, LlamaParse и пр.

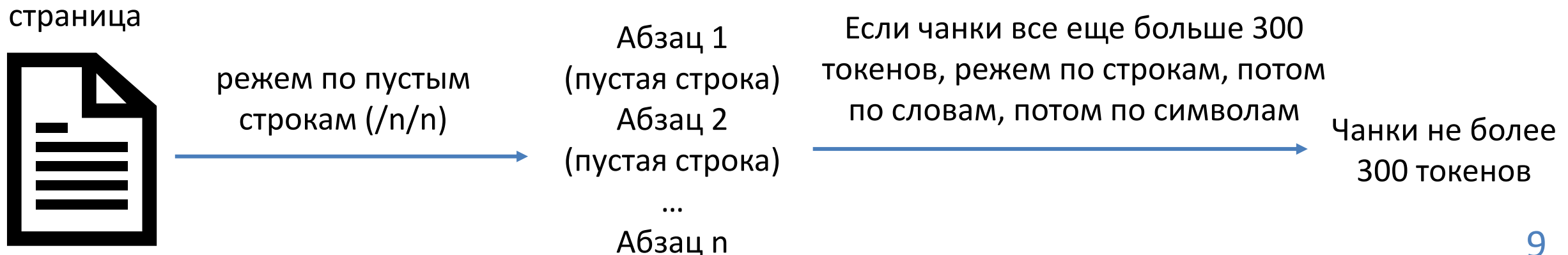
В своем упрощенном примере я использовал pdfplumber.



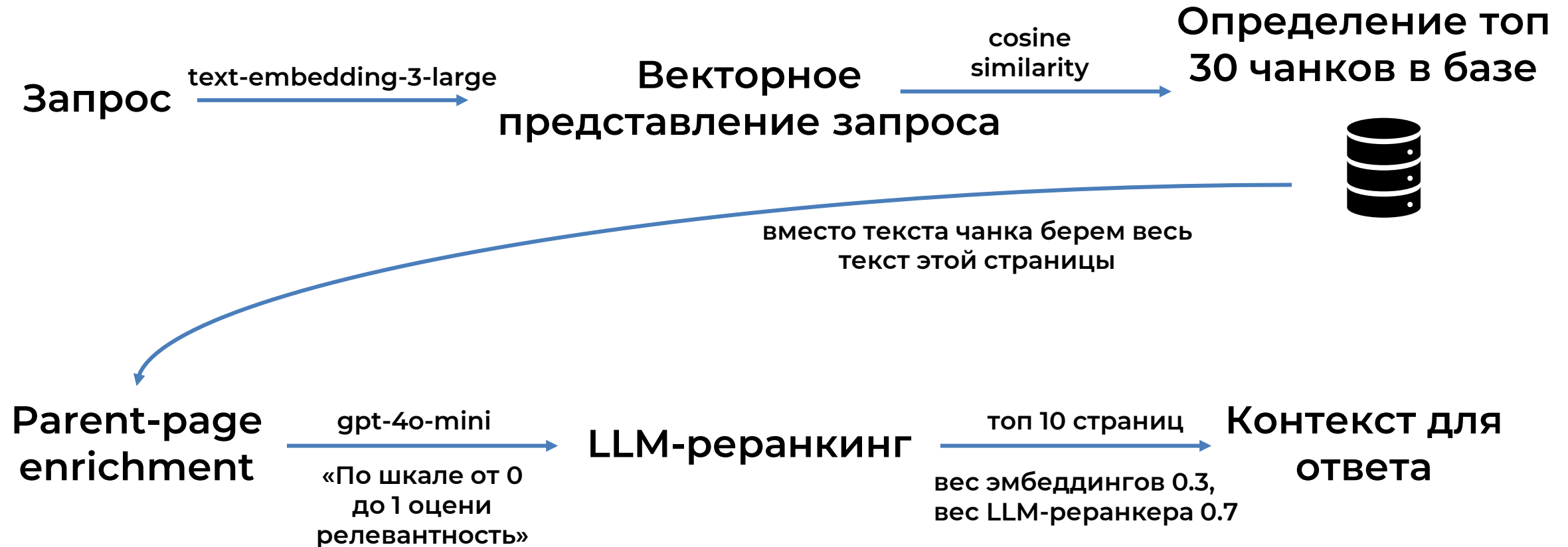
# Ingestion



Как работает рекурсивный md сплиттер:



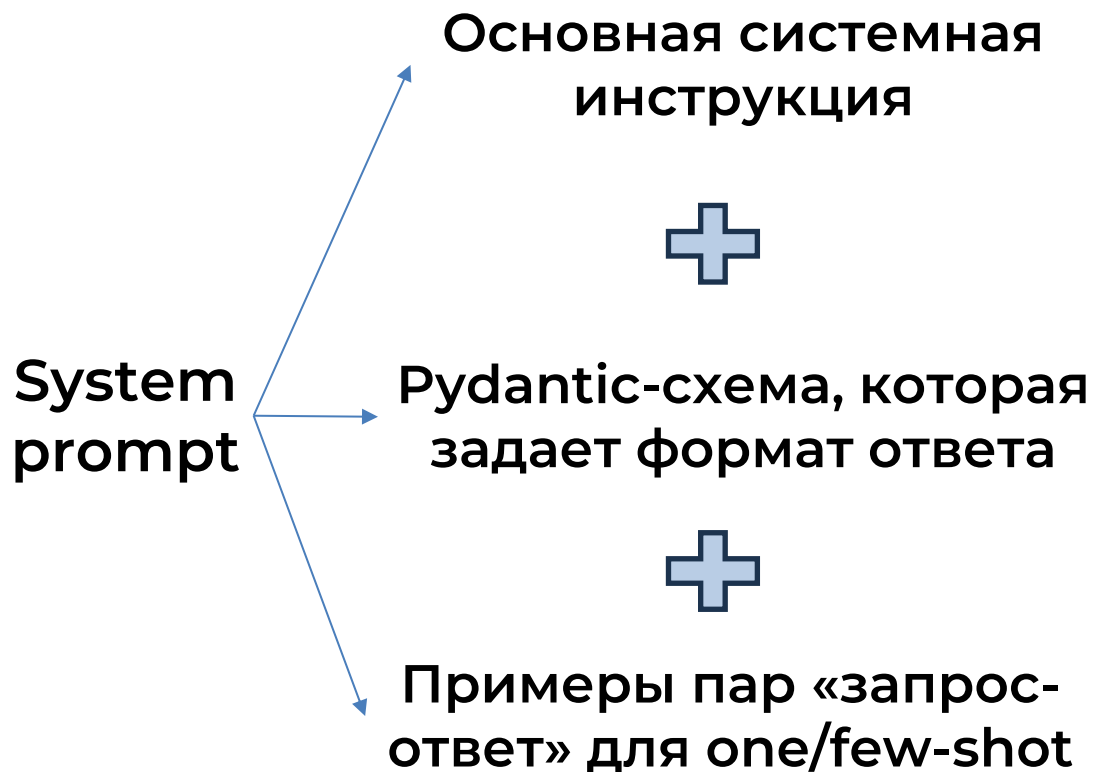
# Retrieval



LLM-переранкинг – перспективная новинка! Обычно используется cross-encoder reranking.

Серебряный призёр вообще построил RAG без эмбеддингов.

# Augmentation



User prompt → Template для вставки контекст и вопроса

Ты — RAG (Retrieval-Augmented Generation) система ответов. Твоя задача — отвечать на заданный вопрос, опираясь ТОЛЬКО на информацию из годового отчёта компании, который предоставляется как релевантные страницы (контекст) после процедуры RAG. Перед финальным ответом внимательно, вслух и по шагам проанализируй вопрос. (и т.д.)

```
step_by_step_analysis: str = Field(description="Подробный пошаговый анализ ответа минимум из 5 шагов и не менее 150 слов.")
str = Field(description="Краткое резюме пошагового анализа. Около 50 слов.") (и т.д.)
```

Пример 1: Какова прибыль Сбера за 2023 год?

```
{"step_by_step_analysis": «...», "reasoning_summary": «...», "relevant_pages": ..., "final_answer": ...}
```

Контекст: (найденный контекст, т.е. отобранные страницы)  
Вопрос: (вопрос)

# Generation

Специфическое для условий данного соревнования решение:

- каждый запрос (кроме multiquery) касался только одного отчета;
- в запросе было явно указано, к отчету какой компании он относится;
- запросы относились к одному из нескольких известных типов.

## 1. Роутинг к нужной БД по имени компании



## 2. Роутинг к нужному шаблону промпта

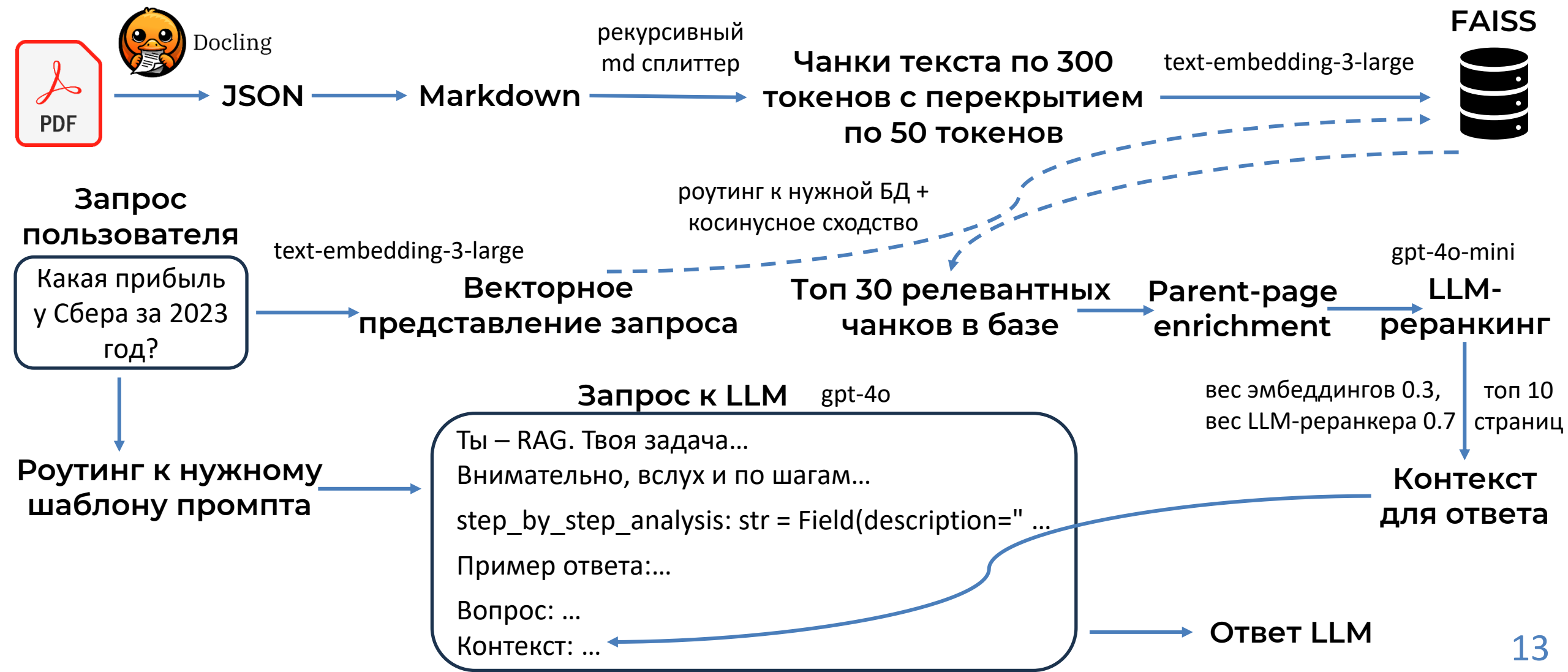


## 3. Роутинг multiquery запросов

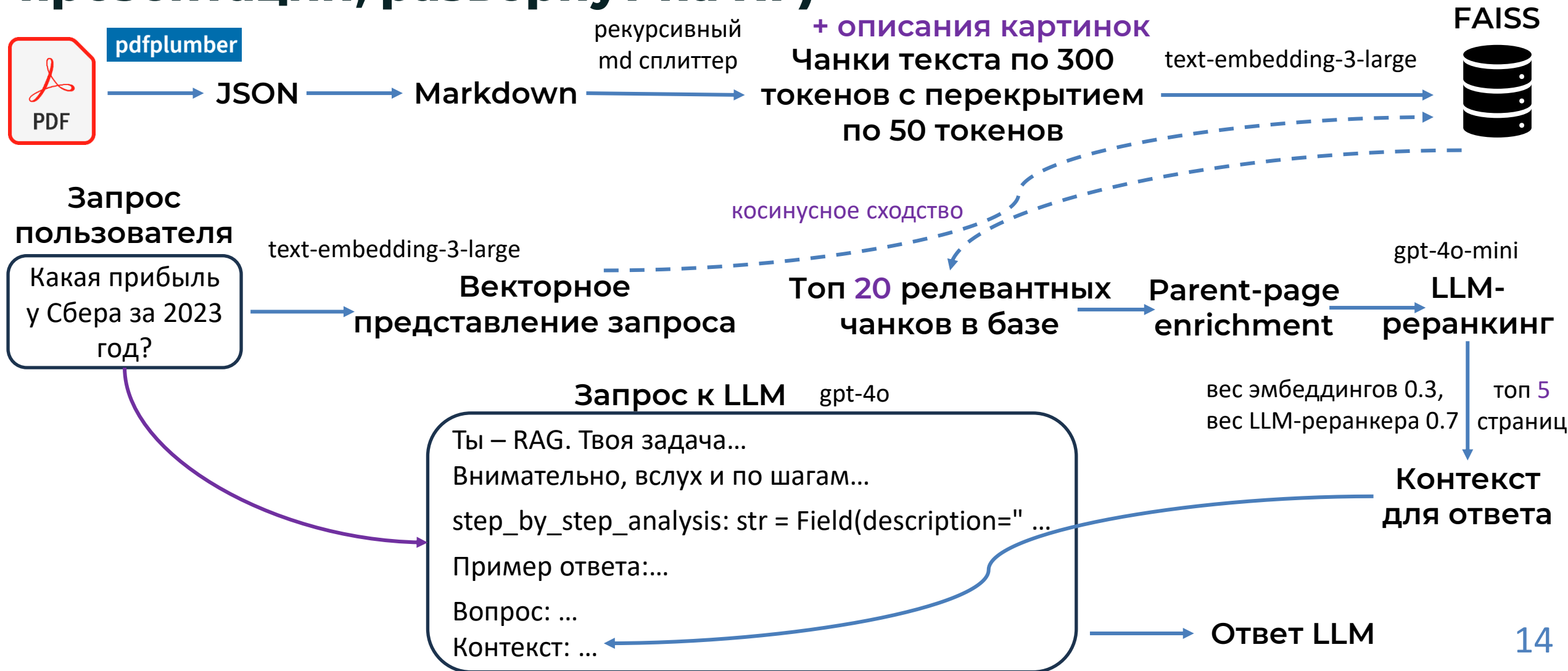




# Ещё раз соберем всё решение вместе



# Упрощенная адаптация (курсовой проект автора презентации, развернут на HF)



# Demo

Задайте вопрос

Ваш вопрос

Каковы основные финансовые показатели Сбера за 2023 год?

Поиск

Ответ системы

✓ **\*\*ФИНАЛЬНЫЙ ОТВЕТ:\*\***

Основные финансовые показатели Сбера за 2023 год включают чистую прибыль в размере 1 511,8 млрд руб., рентабельность капитала (ROE) 25,3%, достаточность капитала 14,0%, и прибыль на акцию 288 руб. Чистый процентный доход составил 2 565 млрд руб., а операционные расходы — 924 млрд руб.

📊 **\*\*Уверенность:\*\*** 90.0%

📖 **\*\*Источники:\*\***

1. Страница 20 (релевантность: 90.0%)  
Ключевые финансовые результаты: Чистая прибыль, рентабельность капитала, достаточность капитала, прибыль на акцию.
2. Страница 186 (релевантность: 80.0%)  
Прибыль на акцию и методы расчета, таблица с финансовыми показателями за 2022 и 2023 годы.
3. Страница 96 (релевантность: 70.0%)  
Анализ отчета о прибылях и убытках, чистая прибыль группы за 2023 год.
4. Страница 95 (релевантность: 70.0%)  
Финансовые показатели: рентабельность активов, рентабельность капитала, чистая процентная маржа.

Источники и статистика

📖 **\*\*Источники:\*\***

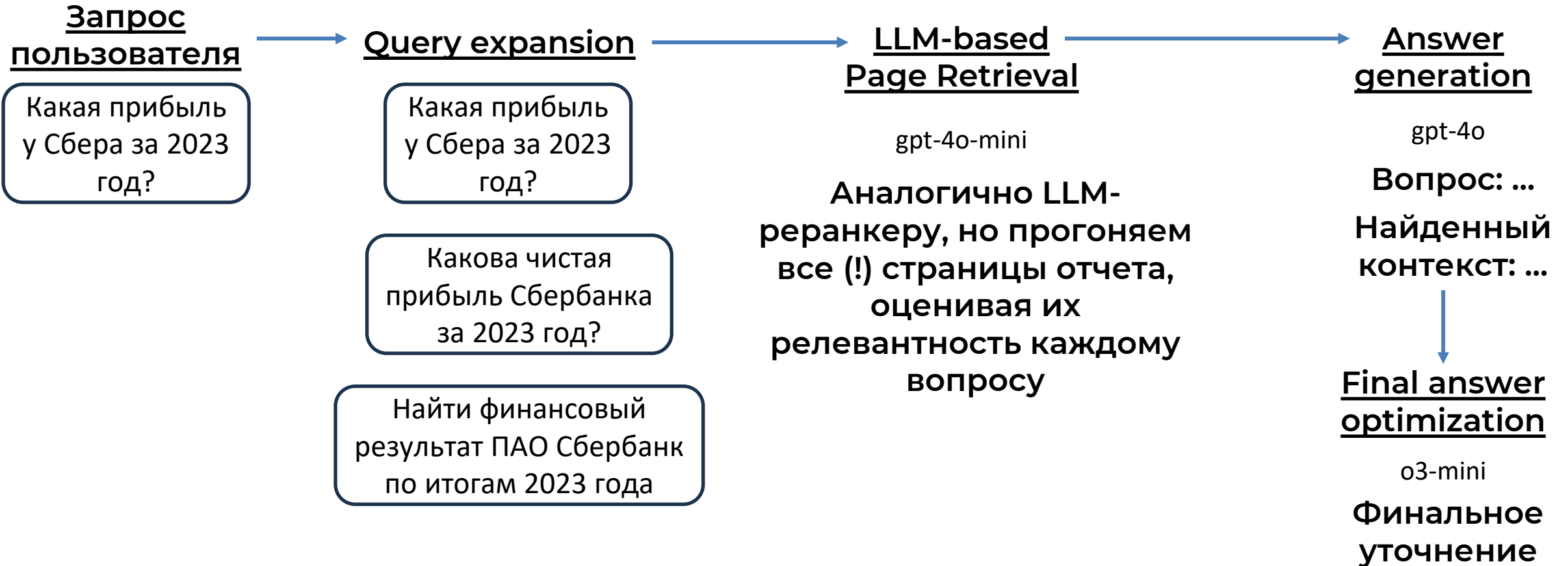
**\*\*1.\*\*** Страница 20 (поиск: 0.761, релевантность: 0.9)

**\*Превью:\*** На изображении представлена инфографика с ключевыми финансовыми результатами Сбера за 2023 год. Показаны чистая прибыль, рентабельность капитала (ROE), достаточность капитала и прибыль на акцию. Графи...

**\*\*2.\*\*** Страница 20 (поиск: 0.744, релевантность: 0.9)

**\*Превью:\*** Ключевые финансовые результаты  
Чистая прибыль,  
млрд руб.  
Собственные средства,  
трлн руб.

# Решение серебряного призера



Стало сенсацией, что можно построить эффективный RAG вообще без эмбеддингов, а на ставших дешевыми и быстрыми, но при этом достаточно качественными для оценки релевантности LLM!



# Выводы

- RAG – одно из основных применений LLM на сегодняшний день.
- Data Science соревнования – один из лучших способов определить наиболее эффективные техники решения задач.
- Победное решение Enterprise RAG Challenge v2 (2025) показало свою эффективность как на больших проприетарных, так и на небольших open source LLM, и победило в обоих треках соревнования – а потому заслуживает изучения и тестирования.
- Качественный Retrieval критически важен.
- LLM-перанкинг, Structured Output, Chain-of-thoughts применялись во многих топ-решениях соревнования.
- Если вы и ваша команда не являетесь профессионалами в RAG или не имеете достаточно ресурсов / времени для реализации и проверки множества различных новомодных RAG-пайплайнов, то адаптация победного решения Enterprise RAG Challenge v2 (2025) является прекрасным обоснованным решением.

# Источники

- ❑ «Как я победил в RAG Challenge: от нуля до SoTA за один конкурс» (<https://habr.com/ru/articles/893356/>)
- ❑ Github-репозиторий победного решения (<https://github.com/IlyaRice/RAG-Challenge-2>)
- ❑ Сайт организатора соревнования Enterprise RAG Challenge (<https://www.timetoact-group.at/en/insights/the-winner-of-the-enterprise-rag-challenge>)
- ❑ Упрощенная адаптация (курсовой проект автора презентации) ([https://huggingface.co/spaces/DmitrySpb/RAG\\_Webinar](https://huggingface.co/spaces/DmitrySpb/RAG_Webinar))
- ❑ Telegram-канал Рината Абдуллина (t.me/llm\_under\_hood)

# Спасибо за внимание!

По любым вопросам:

