

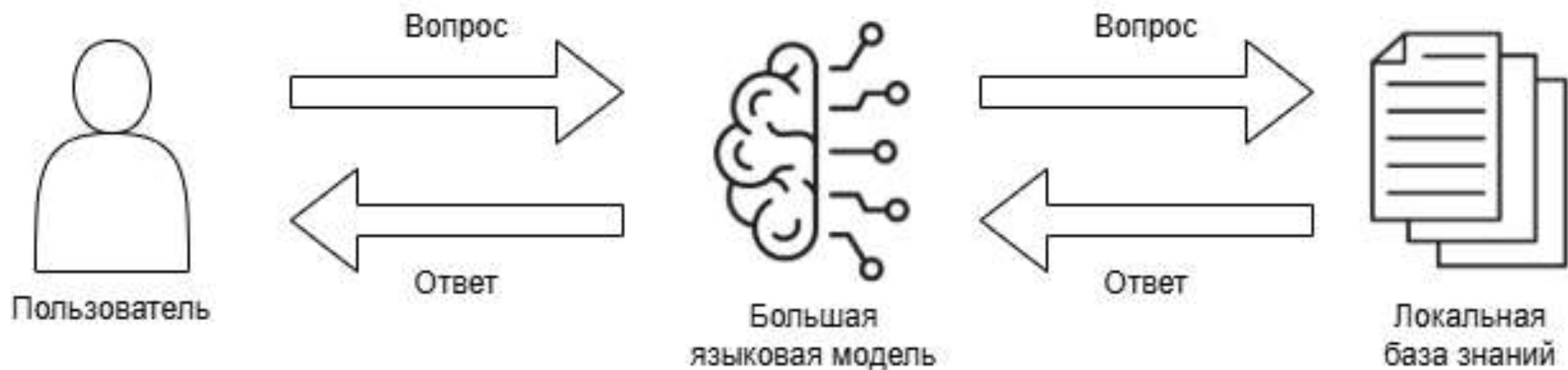
Метод RAG В КОНТЕКСТЕ ПОИСКОВОЙ СИСТЕМЫ сайта (на примере этнографических текстов XVIII века)

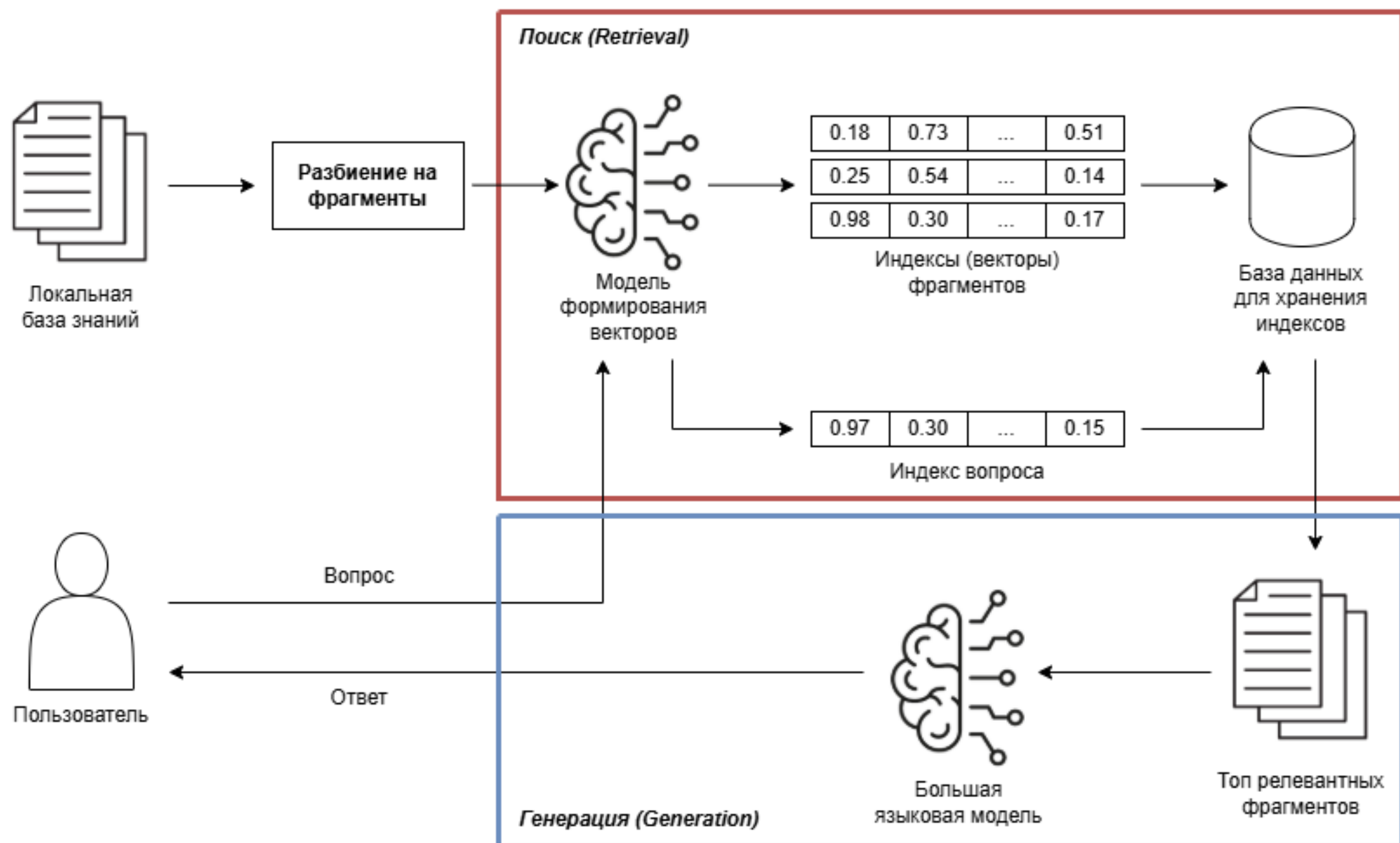
Кира Коваленко
Александр Сергеев
(ЕУСПб)



ЕВРОПЕЙСКИЙ
УНИВЕРСИТЕТ
В САНКТ-ПЕТЕРБУРГЕ

Retrieval-Augmented Generation





УК РФ Статья 247. Нарушение правил обращения экологически опасных веществ и отходов

1. Производство запрещенных видов опасных отходов, транспортировка, хранение, захоронение, использование или иное обращение радиоактивных, бактериологических, химических веществ и отходов с нарушением установленных правил, если эти деяния создали угрозу причинения существенного вреда здоровью человека или окружающей среде, -

наказываются штрафом в размере до двухсот тысяч рублей или в размере заработной платы или иного дохода осужденного за период до восемнадцати месяцев, либо ограничением свободы на срок до двух лет, либо принудительными работами на срок до двух лет, либо лишением свободы на тот же срок.

2. Те же деяния, повлекшие загрязнение, отравление или заражение окружающей среды, причинение вреда здоровью человека либо массовую гибель животных, а равно совершенные в зоне экологического бедствия или в зоне чрезвычайной экологической ситуации, -

наказываются штрафом в размере от ста тысяч до трехсот тысяч рублей или в размере заработной платы или иного дохода осужденного за период от одного года до двух лет, либо принудительными работами на срок до пяти лет, либо лишением свободы на тот же срок.

3. Деяния, предусмотренные частями первой или второй настоящей статьи, повлекшие по неосторожности смерть человека либо массовое заболевание людей, -

наказываются лишением свободы на срок до восьми лет.

Алгоритмы полнотекстовой индексации

- **Матрица термов** – подсчет количества соответствующих слов, встретившихся в тексте
 - Чем чаще слово встречается в запросе или в фрагменте, тем более оно влиятельно
- **TF-IDF** – подсчет количества соответствующих слов с учетом их частотности
 - Часто встречающиеся слова (союзы, местоимения) имеют меньшее влияние в поиске, а узкие редко встречающиеся термины – большее
- **BM25** – функция ранжирования совпадения текстовых фрагментов с запросом на основе TF-IDF

Технологии для поиска

- **LangChain** – программное средство для построения RAG-систем
- **FAISS** – программное средство для эффективного векторного поиска
- **ElasticSearch** и **OpenSearch** – базы данных для хранения и поиска полнотекстовых индексов
- **Milvus** и **Pinecone** – базы данных для хранения и поиска семантических индексов
- Расширения для базы данных **PostgreSQL**:
 - **pgvector** – для семантического поиска
 - **tsvector** – для полнотекстового поиска



Большие языковые модели

- **OpenAI GPT-4** – проприетарная (закрытая, платная) LLM для чат-общения
- **OpenAI o3-mini** – проприетарная (закрытая, платная) LLM для задач, требующих рассуждений
- **Google Gemini** – проприетарная (закрытая, платная) LLM для чат-общения (используется в Google NotebookLM)
- **Яндекс GPT** – проприетарная (закрытая, платная) LLM для чат-общения (используется в Яндекс Нейроэксперт)
- **DeepSeek V3** – открытая LLM для чат-общения
- **DeepSeek R1** – открытая LLM для задач, требующих рассуждений
- **Qwen 3** – открытая LLM гибридного типа, позволяет общаться в чате и вести рассуждения (используется в чат-боте PhytoLex)



Выбор моделей

- [MTEB](#) – оценка моделей для векторных представления (семантического поиска Retrieval, сходства STS) для разных языков

Rank (Box...)	Model	Zero-shot	Memory U...	Number of P...	Embedding D...
1	gemini-embedding-exp-03-07	99%	Unknown	Unknown	3072
2	ling-Embed-Mistral	99%	13563	7B	4096
3	gte-Qwen2-7B-instruct	⚠ NA	29040	7B	3584
4	multilingual-e5-large-instruct	99%	1068	560M	1024
5	SFR-Embedding-Mistral	96%	13563	7B	4096
6	GritLM-7B	99%	13813	7B	4096
7	text-multilingual-embedding-002	99%	Unknown	Unknown	768
8	GritLM-8x7B	99%	89079	57B	4096
9	e5-mistral-7b-instruct	99%	13563	7B	4096
10	Cohere-embed-multilingual-v3.0	⚠ NA	Unknown	Unknown	1024

- [Chatbot Arena](#) – оценка моделей для генерации ответа на основе оценок пользователей по разным задачам и языкам

Rank★ (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI
1	1	Gemini-2.5-Pro-Preview-05-06	1446	+8/-9
2	1	o3-2025-04-16	1413	+8/-7
2	3	ChatGPT-4o-latest (2025-03-26)	1408	+6/-6
3	5	Grok-3-Preview-02-24	1403	+4/-4
4	3	GPT-4.5-Preview	1398	+4/-5
4	5	Gemini-2.5-Flash-Preview-04-17	1394	+7/-7
7	5	DeepSeek-V3-0324	1373	+7/-5
7	5	GPT-4.1-2025-04-14	1366	+7/-8
8	8	DeepSeek-R1	1358	+4/-4
8	15	Gemini-2.0-Flash-001	1355	+4/-3
8	13	Hunyuan-Turbos-20250416	1355	+9/-10
8	5	o4-mini-2025-04-16	1351	+10/-7
10	8	o1-2024-12-17	1350	+4/-4

Примеры RAG-систем

Яндекс Нейроэксперт

<https://expert.ya.ru>

Google NotebookLM

<https://notebooklm.google.com>

Документы 6/25

✓ ozer.txt



✓ petrov_4.txt



✓ sv-pol_1.txt



✓ unknown_3-1.txt



✓ unknown_3-2.txt



✓ moshkov_2.txt



Документы представляют собой фрагменты текстов, связанных с путешествиями и исследованиями различных регионов. Они содержат информацию о планах и маршрутах путешественников, их наблюдениях и намерениях описать увиденные страны и народы. Также упоминаются цели исследований, связанные с изучением природы, обычаев и торговли в разных местах. Документы касаются путешествий по Каспийскому морю, поездок в Персию и другие регионы, а также взаимодействия с местными народами.



Что мы знаем об авторе?

17.05.25 12:53

Ответ на основе 1 источника

Из предоставленных документов можно сделать вывод, что автор:

- обладает глубокими знаниями о состоянии Персии с древних времён до настоящего времени;
- много читает и скупает хорошие книги;
- знает арабский язык и может вести на нём разумные беседы.



Введите вопрос к документам



NotebookLM

Источники

❏

mni_1.txt

❏

❏

mni_3.txt

❏

❏

monten.txt

❏

❏

moshkov_2.txt

☑

❏

ozer.txt

☑

❏

petrov_4.txt

☑

❏

shalfey.txt

❏

❏

sv-pol_1.txt

☑

❏

unknown_3-1.txt

☑

Чат

↻ Обновить

Что мы знаем о Гмелине?

На основе предоставленных источников, мы знаем следующее о Гмелине:

Самуил Готлиб Гмелин был академиком ¹ и испытателем естества ². Он является автором дневных записок о своем путешествии ¹ ...

Его путешествие было поручено Императорской Академией Наук ². Задача состояла в том, чтобы описать страны, по которым он путешествовал, показать их достоинства и недостатки, исследовать особые предметы, примечать домостроительство (экономику), узнать нравы и обычаи народов, словом, собрать все вместе ³. Академия также ...

Его ежегодные наблюдения под видом дневных записок для напечатания ².

↓ Перейти вниз

Введите текст... 6 источников

Студия

Аудиопересказ

Аудиопересказы стали доступны на большем количестве языков. [Подробнее...](#)

🎧 Глубокое обсуждение тем
Двое ведущих

Настроить

Сгенерировать

Заметки

+ Добавить заметку

📖 Методичка

📋 Краткий обзор

🗨 Часто задаваемые вопросы

📅 Хронология

Хронология основных событий:

- **Неизвестная дата:** Создание куч камней возле почитаемого места (вероятно, Богда). Жители, проходя мимо, берут камень у основания и несут его наверх, совершая молитву или оставляя подношения (монеты или кусок одежды).
- **Неизвестная дата:** Появление истории о Богде и озере возле нее.
- **До 11 августа (Год неизвестен):** Автор находится в Аксае, месте проживания казаков, малороссиян и небольшого числа татар.
- **11 августа (Год неизвестен):** Автор покидает Аксай, медленно двигаясь для осмотра трав.
- **До 15 апреля (Год неизвестен):** В Кастинске ходят слухи об открытии селитряных заводов в Урыве.
- **14 апреля (Год неизвестен):** Автор и его спутники оставляют больной скот, закапывают зарезанную корову глубже обычного и поздно ночью возвращаются в Яблонову.
- **15 апреля (Год неизвестен), около 10 утра:** Автор и его спутники отправляются из Яблоновы в Урыв (10 верст), чтобы осмотреть селитряные заводы и другие достопримечательности.

Типичные проблемы

- **Проблемы поиска**

- Пропуск релевантных фрагментов – потеря нужных данных
- Возвращение нерелевантных фрагментов – опора на некорректные данные при ответе
- Устаревшие или смещенные данные – если база знаний содержит устаревшие или предвзятые данные, выводы будут некорректными

- **Проблемы генерации**

- Чересчур строгая опора на факты – если были найдены нерелевантные фрагменты, LLM ответит некорректно, даже если исходно знала ответ
- Некорректная интерпретация – модель может неправильно интерпретировать фрагменты источников

Типичные проблемы

- **Проблемы эксплуатации**

- Вычислительная сложность – работа с большими языковыми моделями требует больших вычислительных ресурсов (памяти, видеоускорителей), но не настолько затратно, как дообучение
- Документы хранятся локально, поиск происходит локально
- Зависимость от используемых технологий – если используются ограниченные или устаревшие технологии, качество результатов может быть низким

- **Этические проблемы**

- Проблема доверия к ответам – LLM слабо поддаются интерпретации, приходится факт-чекать ответы, даже если просим пояснять решения или приводить ссылки
- Смещение и дискриминация – если модели поиска или генерации в процессе обучения приобрела смещение в какой-либо области, ответы могут быть некорректными

NotebookLM

Google заботится о конфиденциальности ваших личных данных и никогда не использует их для обучения NotebookLM.

- Если вы вошли в личный аккаунт Google и оставляете отзывы, специалисты могут проверять ваши запросы, загруженные файлы и ответы ИИ. Это помогает нам устранять неполадки, принимать меры в отношении злоупотреблений и улучшать модель. Не отправляйте информацию, которой вы не хотите делиться с посторонними.
- Если вы пользуетесь Google Workspace или Google Workspace for Education, ваши загруженные файлы, запросы и ответы ИИ в NotebookLM не проверяются специалистами и не используются для обучения ИИ-моделей.

Чат-бот PhytoLex: <http://89.223.99.120:8101>



Чат-бот PhytoLex

Виртуальный ассистент, который прочитал все старинные рукописи о растениях

Привет! Как я могу помочь вам?

Что растет в Москве?

В Московском государстве, как глаголется в старых летописях, растут разные плоды и злаки. Так, "яблокъ, грушъ, вишенъ, дынь, огурцовъ" растет в достатке, а поля "хлѣбородные, всяким землеплодиемъ от Бога обдарена, пшеница, рож, ячмень, проса, овес, гречи́ха и всякие сѣмена, яже суть на потребу человеком, всего родится изобилно" (Фрагмент 1). Окромя того, на мокрых местах близ Москвы можно встретить "растѣніе, названное в Ижорскомъ травникѣ Солнечною росой, а въ простомъ народѣ во многихъ мѣстахъ, особливо около Москвы, подъ именемъ царскихъ очей" (Фрагмент 3).

Благодарим за внимание!