



RAG

ИИ-ассистент для студентов медицинских вузов

Проект выполнили студентки 3-го курса

Кущенко Валерия и Гусева София

Актуальность : кому и зачем это вообще надо ?

Представим себе типичного студента медицинского вуза (*картинка справа*) : вечно уставший, с кучей дедлайнов и недосыпом на все 8 лет обучения.

Как же помочь и немного облегчить их участь ?

— **Сократить время на поиски информации, необходимой для заучивания материалов и выполнения работ.**

Для более тщательного анализа актуальности данной проблемы мы провели опрос среди студентов медицинского вуза.

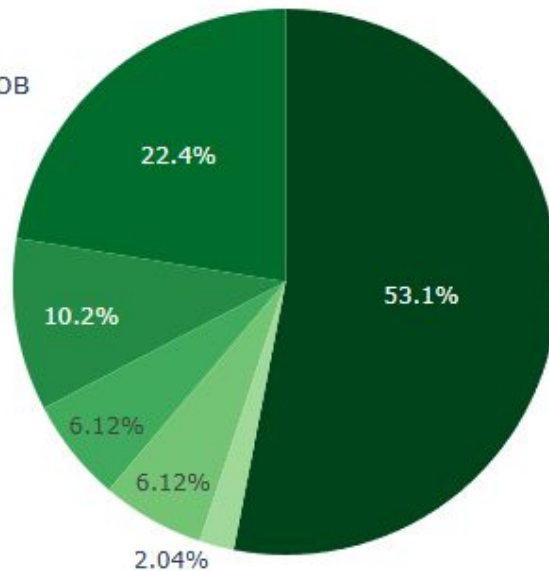


Опрос среди студентов медицинского вуза*

Вопрос:

Когда мне необходимо найти нужную информацию, я иду ...

- Искать её в интернете
- Искать её в книге
- К gpt / deepseek / иным нейронкам
- Всё вместе
- Спрашиваю у однокурсников/одногруппников
- 1-я ступень -- книга, 2-я -- нейронки

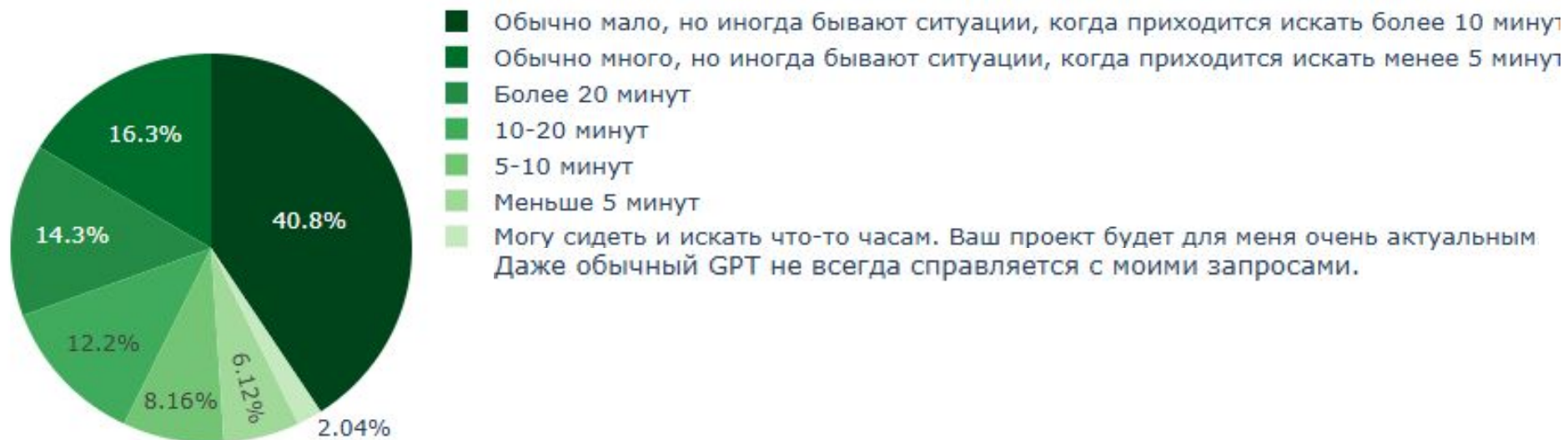


**в опросе приняло участие
около 50 студентов
РНИМУ им. Пирогова.
преимущественно - 3 курс.*

Продолжение опроса

Вопрос:

В среднем на поиск необходимой информации в учебниках я трачу...



Продолжение опроса

Топ-5 самых сложных предметов по мнению студентов (или за что вас могут отчислить)

1 курс	2 курс	3 курс
<ol style="list-style-type: none">1. Анатомия человека2. Гистология3. Общая и биоорганическая химия4. Физика, математика5. Биология	<ol style="list-style-type: none">1. Нормальная физиология2. Биохимия3. Топографическая анатомия и оперативная хирургия4. Анатомия человека5. Пропедевтика внутренних болезней	<ol style="list-style-type: none">1. Топографическая анатомия и оперативная хирургия2. Фармакология3. Патофизиология, клиническая патофизиология4. Патологическая анатомия, клиническая патологическая анатомия5. Пропедевтика детских болезней

Цель и актуальность работы

Актуальность работы:

Как видно из опроса, большинство студентов тратят на поиски информации около 10 (!) минут, и большинство опрошенных первым делом идут искать информацию в интернете, из-за чего могут наткнуться на недостоверные источники.

Цель:

На основе RAG подходов создать ИИ-помощника, который будет по запросу студента искать и генерировать ответ из надежных внутренних источников (книги/учебники, по которым учатся студенты) или внешних (выход в сеть на проверенные сайты).

Задачи

- Проведение опроса среди студентов для выявления актуальности работы ;
- Обзор теории по теме: подходы RAG, и что это вообще такое;
- Сбор данных для обучения модели (надежные учебники/книги);
- Создание рабочей модели RAG;
- Реализация модели в доступном для обычного пользователя виде с помощью Streamlit ;
- Оценка работы ИИ-помощника с помощью метрик: ROUGE-L/ TF-IDF/ оценка ответов модели от студентов.

Анализ теории по теме (краткий)

<i>Retrieval-Augmented Generation for Large Language Models: A Survey.</i>	Эволюция RAG : наивный, расширенный и модульный RAG. Основные стадии RAG – извлечение, генерация и аугментация.
<i>Benchmarking Large Language Models in Retrieval-Augmented Generation</i>	Нет строгой оценки влияния RAG на работу разных LLM. Сравнение работы 6 разных LLM по четырем критериям: устойчивость к шуму, отказ от ненужной информации, интеграция информации (может ли отвечать на сложные вопросы), устойчивость к ложным данным.
<i>A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions</i>	Описание общей архитектуры. Выявление существующих проблем в системах RAG, обзор областей для дальнейшего развития.
<i>Corrective Retrieval Augmented Generation</i>	RAG частично решает проблему галлюцинаций LLM, но если поиск неудачен, то генерация страдает. Авторы предлагают для решения этой проблемы оценивать качество извлеченной информации и на ее основе, либо использовать внутренние источники, либо выходить во внешний поиск (Интернет).

Retrieval-Augmented Generation for Large Language Models: A Survey.

Наивный RAG:

- индексация ;
- извлечение ;
- генерация.

*Проблемы с извлечением данных, трудностью генерации и препятствие к дополнению информации.

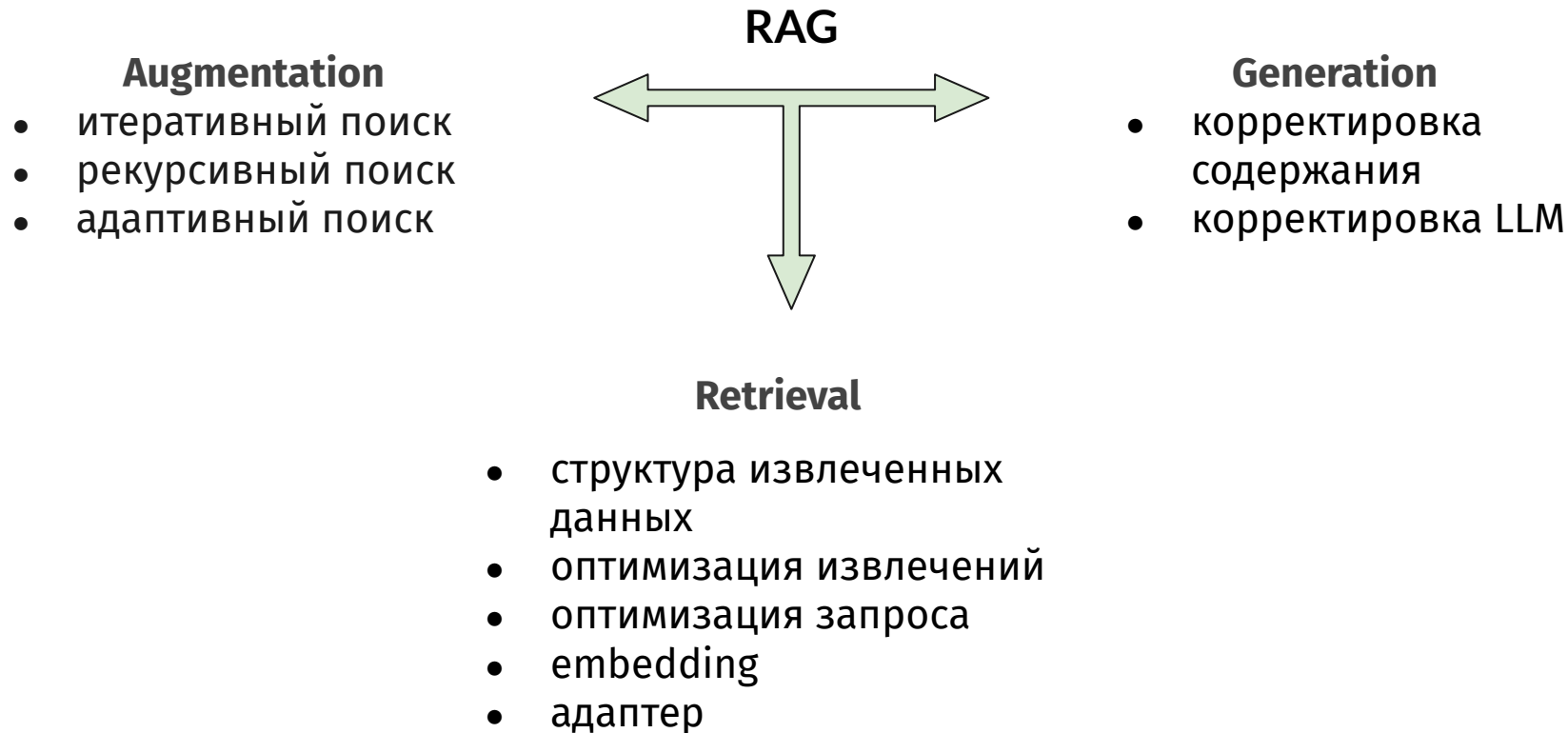
Расширенный RAG:

- улучшение техник индексации с помощью подхода со скользящим окном;
- до извлечения – оптимизация индексной структуры и исходного запроса;
- после извлечения – повторная сортировка фрагментов и сжатие контекста.

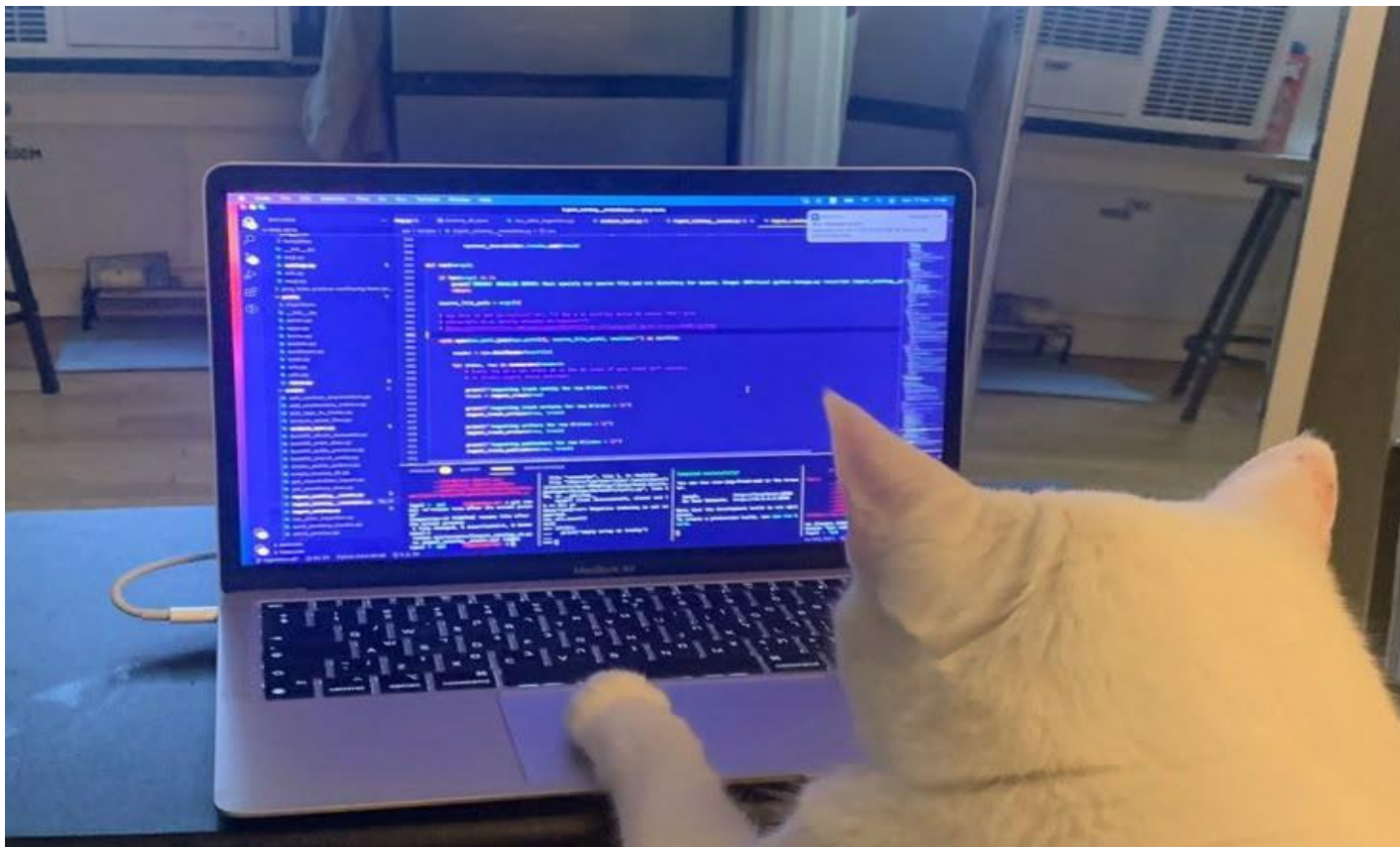
Модульный RAG:

- добавление новых модулей (поисковой, слияния, памяти, предсказания, маршрутизации, адаптации);
- добавление новых паттернов (уточнение запросов на извлечение через модуль rewrite).

Retrieval-Augmented Generation for Large Language Models: A Survey.



Пора перейти к нашей работе....



Спецификация модели

llama3.2

ollama run llama3.2



↓ 17M Downloads ⌚ Updated 7 months ago

Meta's Llama 3.2 goes small with 1B and 3B models.

[tools](#) [1b](#) [3b](#)

Sizes

3B parameters (default)



The 3B model outperforms the Gemma 2 2.6B and Phi 3.5-mini models on tasks such as:

- Following instructions
- Summarization
- Prompt rewriting
- Tool use

ollama run llama3.2

LLAMA3.2

- Занимает не так много памяти
- Подходит для использования инструкций
- Поддерживает русский язык

 [intfloat/multilingual-e5-base](#) 

♡ like 270



Sentence Similarity



sentence-transformers



PyTorch



ONNX





EMBEDDINGS from
intfloat/multilingual-e5-base

Tavily for WebSearch

Pages / Overview

Overview

Operational



CURRENT PLAN

Manage Plan

Researcher

API Usage ⓘ





Plan

0 / 1000 Credits

☐ Pay as you go ⓘ

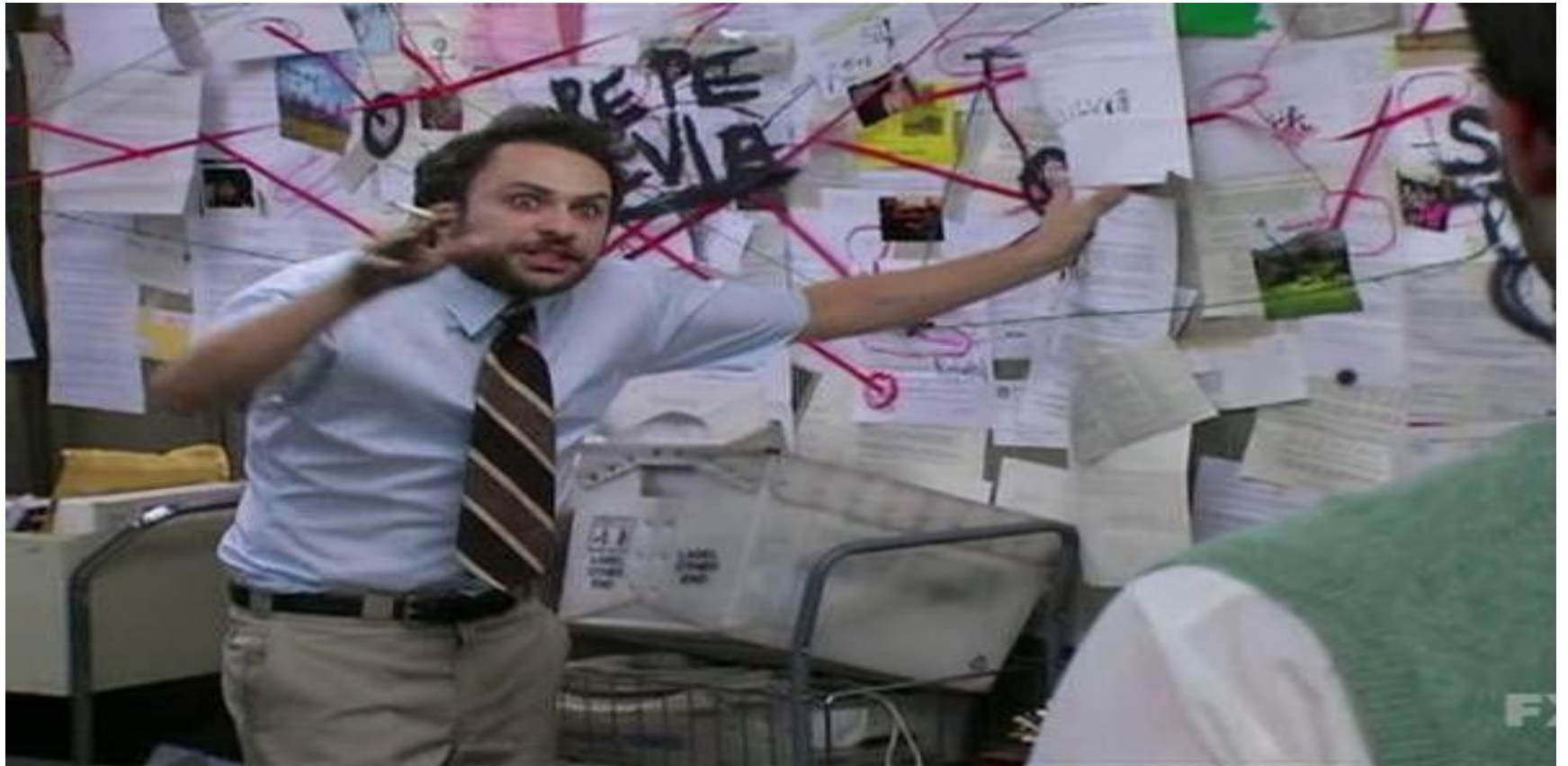
API Keys +

The key is used to authenticate your requests to the [Research API](#). To learn more, see the [documentation](#) page.

NAME	TYPE	USAGE	KEY	OPTIONS
default	dev	0	tvly-dev-*****	   

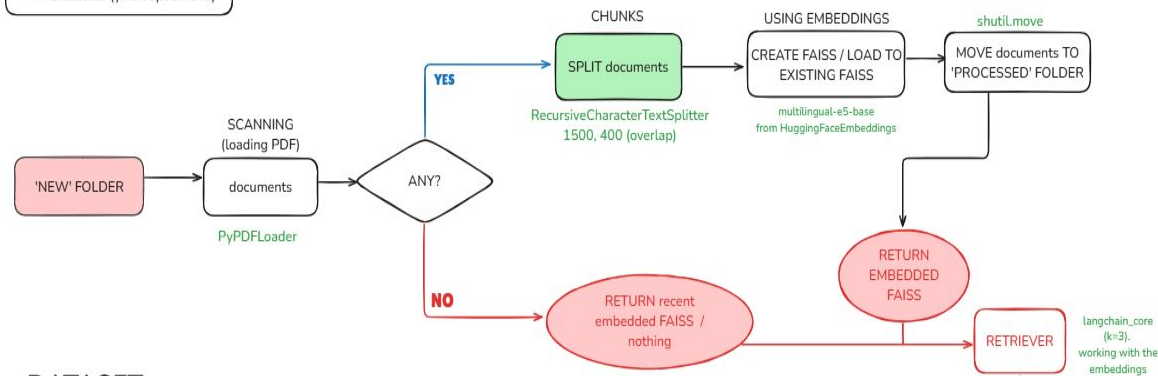


Архитектура



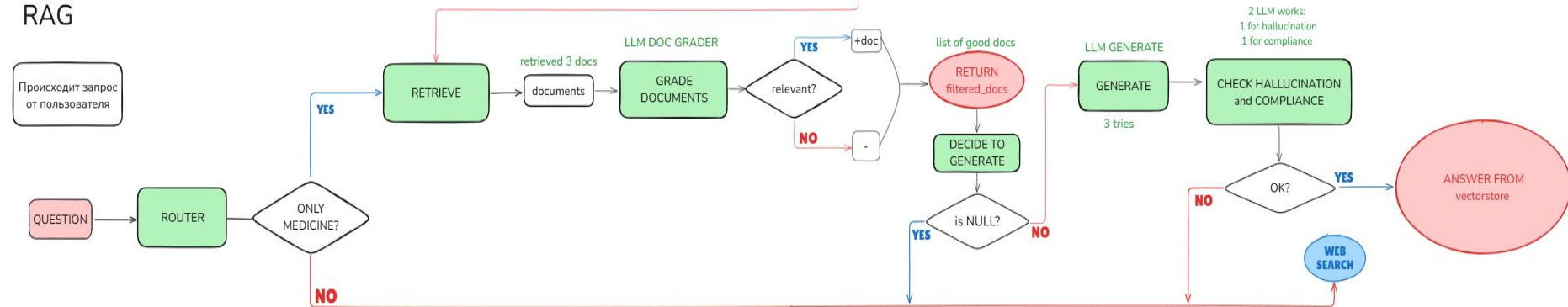
Архитектура

Существует 2 папки:
- NEW (для необработанной лит.)
- PROCESSED (уже обработана)



DATASET

RAG



Данные для обучения: откуда брали и на чем учили

1. Литература, на которой обучалась модель:

(С полным списком литературы можно ознакомиться в разделе About Model на нашем сайте)

- “Введение в молекулярную биологию от клеток к атомам” - Э. Рис, М. Стернберг;
- “Биологическая химия” - Е.С. Северин, Т.Л. Алейникова, Е.В. Осипов и С.А. Силаева;
- “Основы патологии: Учебное пособие для студентов медицинских училищ и колледжей” - И.В. Ремизов, В.А. Дорошенко;
- “Общая хирургия” - Гостищев В.К;
- “Пропедевтика детских болезней с уходом за детьми” - Т.В. Капитан;
- и другие.

2. Выход во внешние источники с помощью ресурса Tavily AI.

3. Возможность самостоятельной загрузки файлов со стороны студента.

**Вся литература была подобрана на основе её наличия в программе обучения у разных (признанных) вузов*

Пример работы модели



Что такое рак легкого?



Ассистент: Рак легкого развивается из эпителия бронхов разного калибра и очень редко из альвеолярного эпителия. В подавляющем числе случаев рак легкого является бронхогенным раком легкого, а в редких случаях - альвеолярным раком. Рак легкого различают по локализации: прикорневой, периферический и смешанный. По микроскопической картине рак легкого можно разделить на плоскоклеточный, железистый и недифференцированный типы.

Рак легкого.

В подавляющем числе случаев рак легкого развивается из эпителия бронхов разного калибра (**бронхогенный рак легкого**) и очень редко из альвеолярного эпителия.

В зависимости от локализации различают **прикорневой (центральный) рак**, исходящий из главного, долевого или начальной части сегментарного бронха, **периферический рак**, исходящий из периферического отдела сегментарного бронха и его ветвей, а также альвеолярного эпителия, и **смешанный рак**.

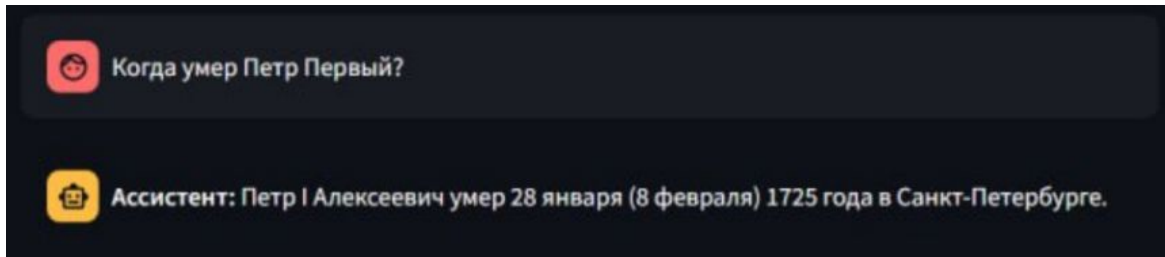
Далее приведен пример работы модели.

Ей был задан вопрос из учебника *“ОСНОВЫ ПАТОЛОГИИ Учебное пособие для студентов медицинских училищ и колледжей”* - эталонный текст приведен на второй фотографии.

Как можно заметить, модель легко справилась с вопросом и дала качественный ответ.

Пример работы модели (WebSearch)

Также приведем пример работы модели с не релевантным текстом. Все учебники содержат в себе информацию ТОЛЬКО о медицине, поэтому было решено задать исторический вопрос.



Модель справилась с заданием и ушла в WebSearch с помощью Tavily - это отражено в логах.

```
2025-05-08T02:01:06.536381+0300 DEBUG {'question': 'Когда умер Петр Первый?', 'web_search': 'Yes',  
2025-05-08T02:01:06.583244+0300 DEBUG ---ПОИСК в ИНТЕРНЕТЕ---  
2025-05-08T02:01:06.583244+0300 DEBUG ---ПОИСК в ИНТЕРНЕТЕ---  
2025-05-08T02:01:06.583244+0300 DEBUG ---ПОИСК в ИНТЕРНЕТЕ---
```

Демонстрация работы модели с новой литературой

MedBro

localhost:8501

Deploy

Agent

About project

Statistics


About model


Узнайте о проекте чуть больше

При желании загрузите свой PDF-файл

Drag and drop file here
Limit 200MB per file • PDF

Browse files



 **ИИ-ассистент для студентов
медицинских университетов**

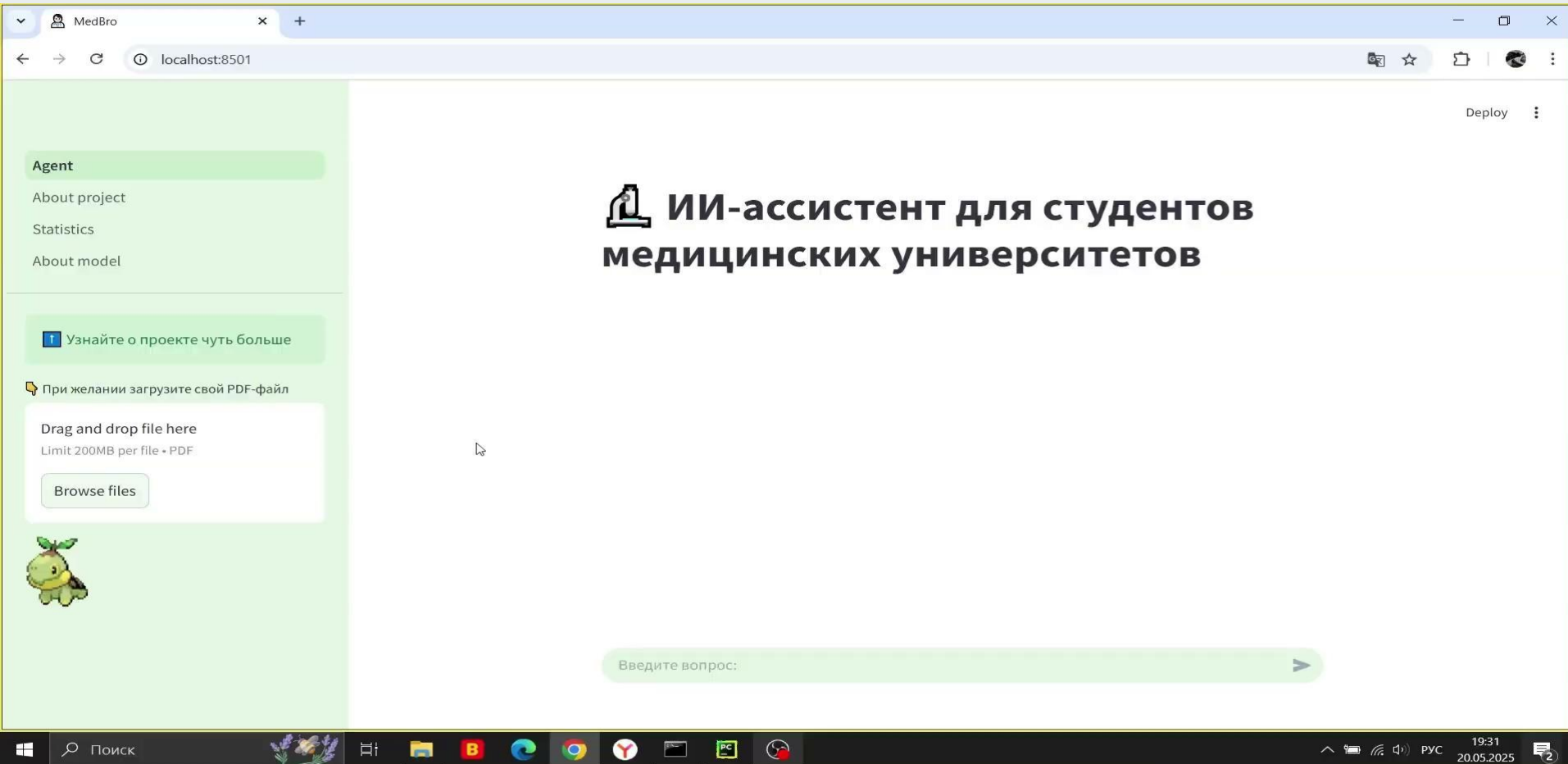
Введите вопрос:

Поиск



18:49
20.05.2025

Как устроен сайт



Оценка работы модели: метрики качества

Модели было задано **15 вопросов** по медицинской тематике. Позже ответы модели сравнивались с “эталонными” ответами на те же вопросы, но уже из учебников.

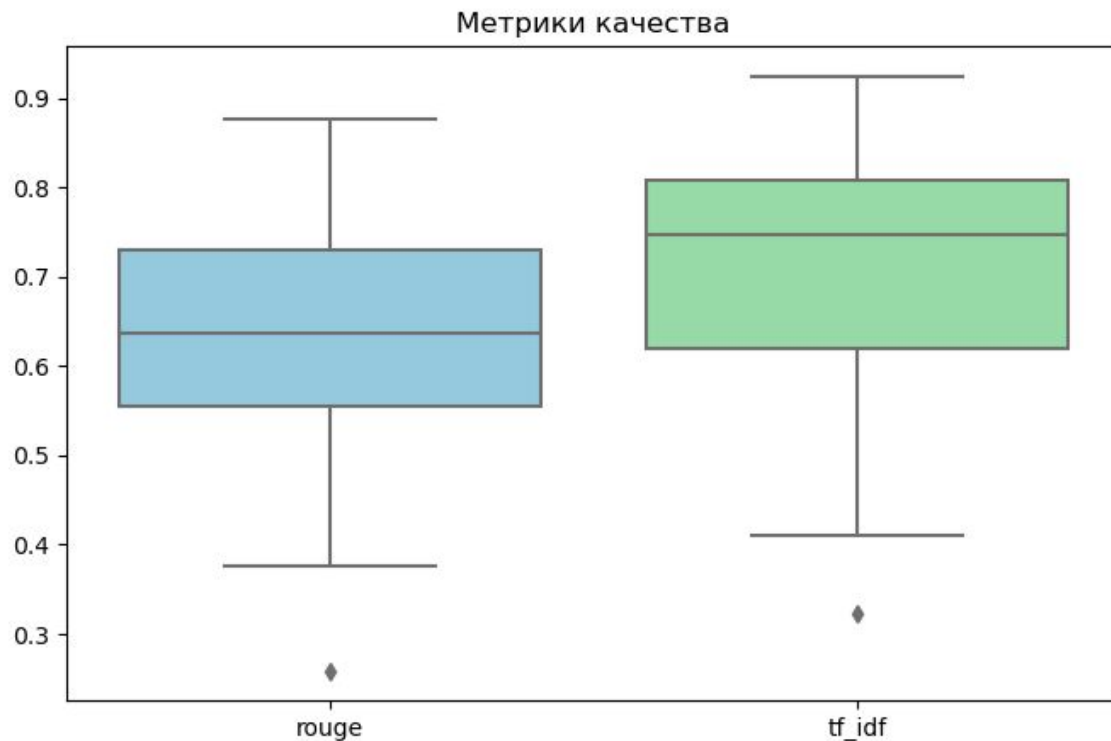
Для сравнения ответов были использованы **три метрики качества**.

Метрики :

- ROUGE-L (F1-score)
- TF-IDF + косинусное расстояние
- Кастомная метрика : оценка ответов модели от студентов-отличников медицинского вуза.

Средние значения для каждой метрики		
ROUGE-L (F1-score)	TF-IDF + косинусное расстояние	Кастомная метрика
0.6227	0.6912	0.9467

Оценка работы модели: TF-IDF и ROUGE-L



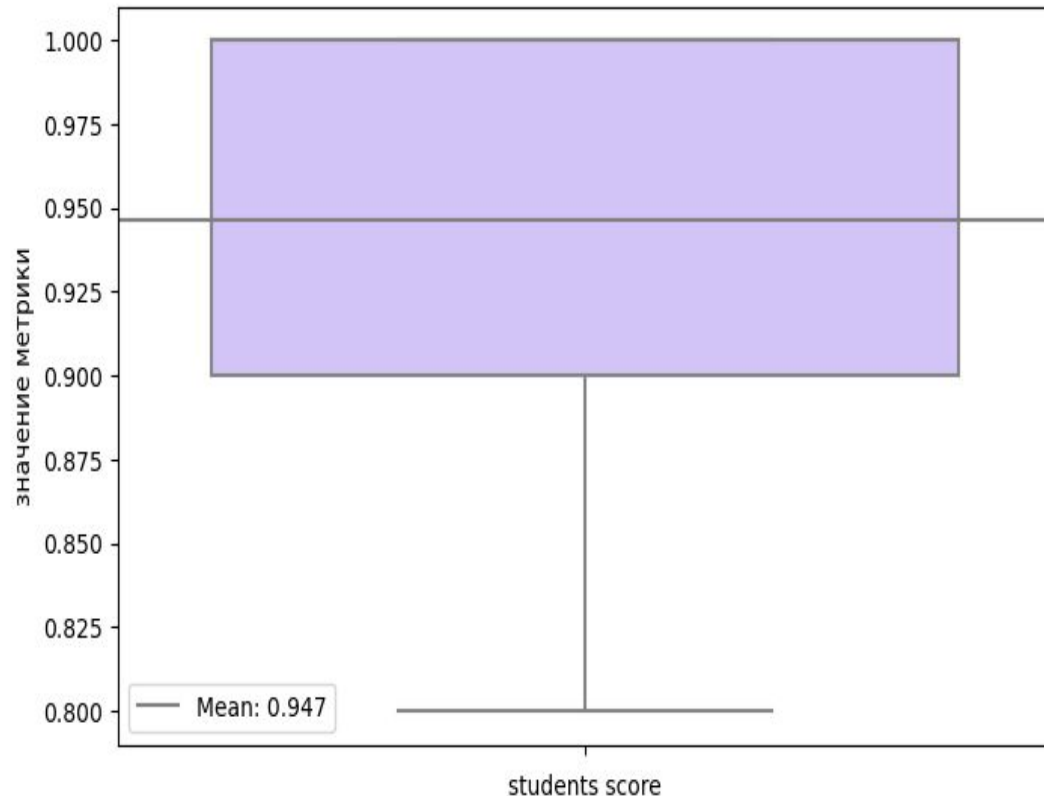
	ROUGE-L	TF-IDF
max	0.88	0.92
median	0.64	0.75
min	0.26	0.32
mean	0.62	0.69

Оценка работы модели: кастомная метрика

Мы попросили 5х студентов-отличников оценить ответы модели на 15 разных вопросов из медицинской сферы, предоставив им при этом “эталонный” вариант ответа из книги.

9 из 15 ответов модели по мнению всех 5х студентов по содержанию полностью соответствуют эталонному ответу из книги.

Среднее значение метрики: 0.95



Возможности для развития проекта

- Подбор литературы, на которой обучается модель, совместно с экспертом в медицинской сфере;
- Добавление выпадающего списка в панели для выбора специального предмета. Выбор предмета “Анатомия” гарантировал бы базу знаний, состоящую из анатомических книг.
- Работа не только с книгами формата PDF, но и с отсканированными документами (конспектами);
- Оценка совпадения загруженной литературы с медицинской тематикой.
- Вывод ссылки на источник ответа модели.

Этика

Данный проект создан с целью помочь **СТУДЕНТАМ медицинских вузов** в их обучении. Модель не способна навредить человеку – в крайнем случае студент может проконсультироваться с преподавателем, и он укажет ему на ошибки. Тем не менее, риск ошибки модели минимален из-за проверки на галлюцинации и опоры на текст.

Ссылки на источники, которые используются в работе

- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, Haofen Wang. [Retrieval-Augmented Generation for Large Language Models: A Survey](#). — 2024-03-27.
- Shailja Gupta, Rajesh Ranjan, Surya Narayan Singh. [A Comprehensive Survey of Retrieval-Augmented Generation \(RAG\): Evolution, Current Landscape and Future Directions](#). — 2024-10-03.
- Jiawei Chen, Hongyu Lin, Xianpei Han, Le Sun. [Benchmarking Large Language Models in Retrieval-Augmented Generation](#). — 2023-12-20
- Corrective Retrieval Augmented Generation Shi-Qi Yan^{1*}, Jia-Chen Gu^{2*}, Yun Zhu³, Zhen-Hua Ling¹
- https://github.com/kvoloshenko/Local_RAG_Agent_01
- https://langchain-ai.github.io/langgraph/tutorials/rag/langgraph_adaptive_rag_local/