

RAG-система с LLMs

Краснобаева Вера,
Семенюк Полина,
Васяткина Анастасия



Цели и задачи

Цель проекта:

- На основе представленного датасета и LLM реализовать RAG систему
- RAG (Retrieval-Augmented Generation) — это система, объединяющая поиск и генерацию текста

Задачи:

- Реализовать вариант модели без retrieval части (базовая модель)
- Реализовать вариант с retrieval частью (улучшенная модель) с использованием векторной базы данных
- Основная гипотеза: использование векторной базы данных должно улучшить качество модели






Данные

- **Датасет** для оценки RAG систем, размер 50.4k
<https://huggingface.co/datasets/bearberry/sberquadqa>
- **Содержит:**
 - Вопрос
 - Правильный ответ на него
 - Тот же ответ, но все слова в нем приведены к начальной форме
 - Контекст = текст, разделенный на чанки, из которого нужно достать ответ на вопрос
- Проверим работу модели на 50 примерах в режиме zero-shot

Split (1)

train · 50.4k rows

Search this dataset

id	question	answers	normalized_answers	context	metadata
string · lengths	string · lengths	sequence · lengths	sequence · lengths	list · lengths	dict
 10-11 <0.1%	 8-75 72.6%	 1-2 98.9%	 1-2 98.9%	 1-6 62.6%	
0_sberquad	чем представлены органические остатки?	["известковыми выделениями сине-...	["известковый выделение синий...	[{ "chunk": "В протерозойских...	{ "is_answerable":...
1_sberquad	что найдено в кремнистых сланцах...	["нитевидные водоросли, грибные...	["нитевидный водоросль грибной...	[{ "chunk": "В протерозойских...	{ "is_answerable":...
2_sberquad	что встречается в протерозойских отложениях?	["органические остатки"]	["органический остаток"]	[{ "chunk": "В протерозойских отложениях органические остатки встречаются намного чаще, чем в архейских.", "is_relevant": true }, { "chunk": "Они представлены известковыми выделениями сине-зеленых водорослей, ходами червей, остатками кишечнорастворимых.", "is_relevant": false }, { "chunk": "Кроме известковых водорослей, к числу древнейших растительных остатков	{ "is_answerable": true, "tag": null }

Предобработка данных (для подсчета метрик)

Нормализация ответов

- Лемматизация
- Стемминг
- Синонимизация

Такая предобработка позволяет смягчить метрику.

Исходная модель для генерации ответа

- RefalMachine/RuadaptQwen2.5-1.5B-instruct + Инструктивная версия адаптированной на русский язык модели Qwen2.5-1.5B. В модели был заменен токенизатор, затем произведено дообучение (Continued pretraining) на русскоязычном корпусе
- Qwen — это семейство больших языковых моделей (LLM), разработанных компанией Alibaba Cloud. Они ориентированы на широкий спектр задач, включая генерацию текста, кодирование, диалоговые системы и RAG-приложения. Плюсы: многоязычны, оптимизированы под RAG

Реализация части с retrieval

Qdrant — это векторная база данных с открытым исходным кодом, предназначенная для поиска ближайших соседей. Она часто используется в RAG-системах для хранения и поиска эмбеддингов текста, изображений и других данных.

- Загружаем документы → Разбиваем на куски → Создаём эмбеддинги → Сохраняем в Qdrant.
- Получаем запрос пользователя → Преобразуем в вектор → Ищем похожие вектора в Qdrant.
- Берём найденные фрагменты → Отправляем в LLM → Генерируем осмысленный ответ.

Эксперименты по улучшению качества

1. Подбор параметров:

- `window_size` – количество релевантных чанков
- `top_key` – аналог `window_size` для модели с извлечением
- `num_beams` – количество лучей
- `max_new_tokens` – максимальная длина ответа модели

2. Повтор релевантного чанка несколько раз в промте

3. Подбор разных моделей

- Для генерации ответа
- Для векторизации

Варьируем LLM

- **RefalMachine/RuadaptQwen2.5-1.5B-instruct (1.53B)**
- Vikhrmodels/Vikhr-Gemma-2B-instruct (2.61B)
- Vikhrmodels/QVikhr-2.5-1.5B-Instruct-r (1.54B)
- Vikhrmodels/Vikhr-Llama-3.2-1B-Instruct (1.24B)
- MTSAIR/Cotype-Nano (1.54B)

Результаты для разных моделей

rep, max_new_tokens, num_beams, window_size		False, 15, None, 10	True, 15, 5, 10	False, 15, 5, 10	True, 15, 5, 10	True, 20, 5, 10	False, 20, 5, 10
RefalMachine/RuadaptQwen2.5-1.5B-instruct 1.53B	morph	61,974453	71,168203	63,538872	67,412021	68,309890	66,249068
	stem	60,772072	69,955749	63,231180	66,770995	67,305228	64,536613
	vectors	66,829268	77,073171	69,287469	72,371638	71,770335	68,527919
Vikhrmodels/Vikhr-Gemma-2B-instruct 2.61B	morph	52,908956	72,437202	63,441575	69,866228	71,719408	67,385553
	stem	50,432766	69,488484	62,341575	68,320441	70,173621	66,385553
	vectors	62,053571	77,486911	67,120181	75,062972	76,070529	72,235872
Vikhrmodels/QVikhr-2.5-1.5B-Instruct-r 1.54B	morph	51,590306	56,028255	49,520973	55,041353	56,568471	54,877620
	stem	50,923639	54,049234	47,875286	52,617888	54,710329	53,873735
	vectors	53,580902	64,566929	57,364341	64,583333	64,319249	58,494624
Vikhrmodels/Vikhr-Llama-3.2-1B-Instruct 1.24B	morph	44,987446	48,706349	47,416511	48,869048	50,003602	48,899328
	stem	44,054113	47,173016	47,216511	47,869048	48,781380	48,121551
	vectors	56,818182	55,737705	55,056180	51,752022	54,271357	58,942065
MTSAIR/Cotype-Nano 1.54B	morph	46,375523	61,549756	47,580510	62,689317	66,599151	59,868476
	stem	45,631796	60,570735	45,712600	61,710296	65,231630	58,809035
	vectors	52,132701	66,165414	53,588517	66,498741	69,284065	60,485651

Варьируем модель для получения векторов

- Модели:
 1. RefalMachine/RuadaptQwen2.5-1.5B-instruct (1.53B)
 2. Vikhrmodels/Vikhr-Gemma-2B-instruct (2.61B)

	No retrieval			Retrieval			Retrieval: ai-forever/sbert_large_nlu_ru		
	False, 15, None, 10	True, 15, 5, 10	False, 15, 5, 10	True, 15, 5, 10	True, 20, 5, 10	False, 20, 5, 10	True, 15, 5, 10	True, 20, 5, 10	False, 20, 5, 10
morph	61,974453	71,168203	63,538872	67,412021	68,309890	66,249068	51,837745	52,241531	66,358154
stem	60,772072	69,955749	63,231180	66,770995	67,305228	64,536613	50,125291	50,529076	65,342770
vectors	66,829268	77,073171	69,287469	72,371638	71,770335	68,527919	63,569682	61,904762	68,181818
morph	52,908956	72,437202	63,441575	69,866228	71,719408	67,385553	60,004024	60,440965	69,621984
stem	50,432766	69,488484	62,341575	68,320441	70,173621	66,385553	57,518554	57,955495	68,288650
vectors	62,053571	77,486911	67,120181	75,062972	76,070529	72,235872	65,664160	67,167920	75,810474

Анализ результатов

- Используем f-метрику для всех моделей
1. По качеству результатов с моделью, которую мы выбрали исходно как основную (*RefalMachine/RuadaptQwen2.5-1.5B-instruct (1.53B)*), сравнима только модель *Vikhrmodels/Vikhr-Gemma-2B-instruct (2.61B)*, но при этом она значительно большего размера, что влияет на производительность.
 2. Альтернативная модель для получения векторов (*ai-forever/sbert_large_nlu_ru*) дает худшие результаты, чем наша исходная модель для векторизации (*intfloat/multilingual-e5-large*).
 3. Использование базы данных (retrieval) не дает улучшения результатов. Но с ним модель показывает более стабильные результаты.

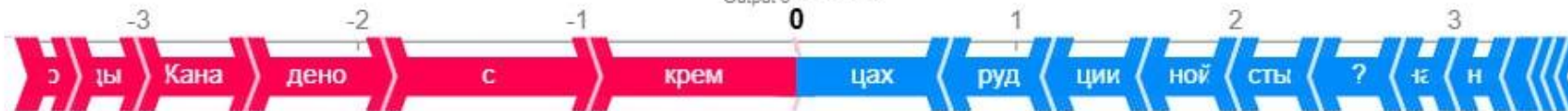
Интерпретируемость

- Добавляем explainability для модели Vikhrmodels/Vikhr-Gemma-2B-instruct (2.61B), показавшей лучший результат.
- Используем библиотеку shap метод explain
- Сравним результат для трех наборов параметров

js

outputs
Output 0

f_base(values)
Output 0

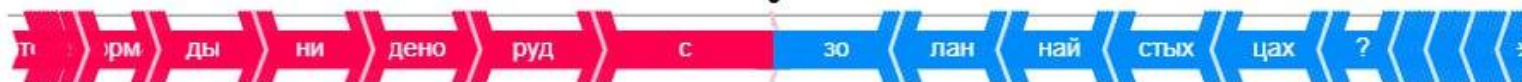


inputs
что найдено в кремнистых сланцах железорудной формации Канады?

js

outputs
Output 0

f_base(values)
Output 0

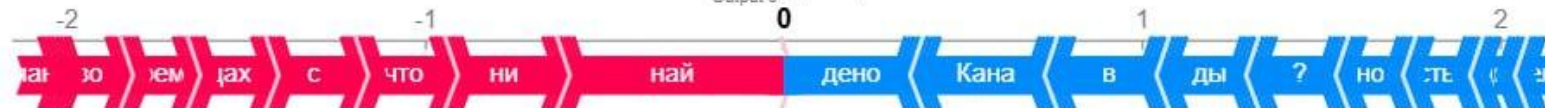


inputs
что найдено в кремнистых сланцах железорудной формации Канады?

js

outputs
Output 0

f_base(values)
Output 0



inputs
что найдено в кремнистых сланцах железорудной формации Канады?

Спасибо за внимание!