

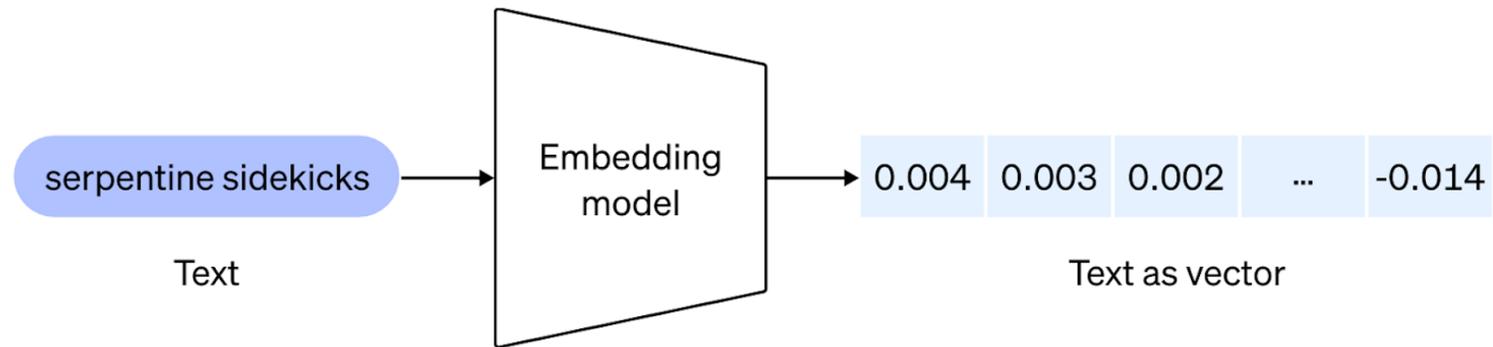
GigaEmbeddings

Колодин Егор, Хомич Дарья

**GIGA
CHAT**



Зачем нужны embedding модели



Отображаем текст в вектор для использования в последующих задачах классификации, ранжирования, поиска и т. п.



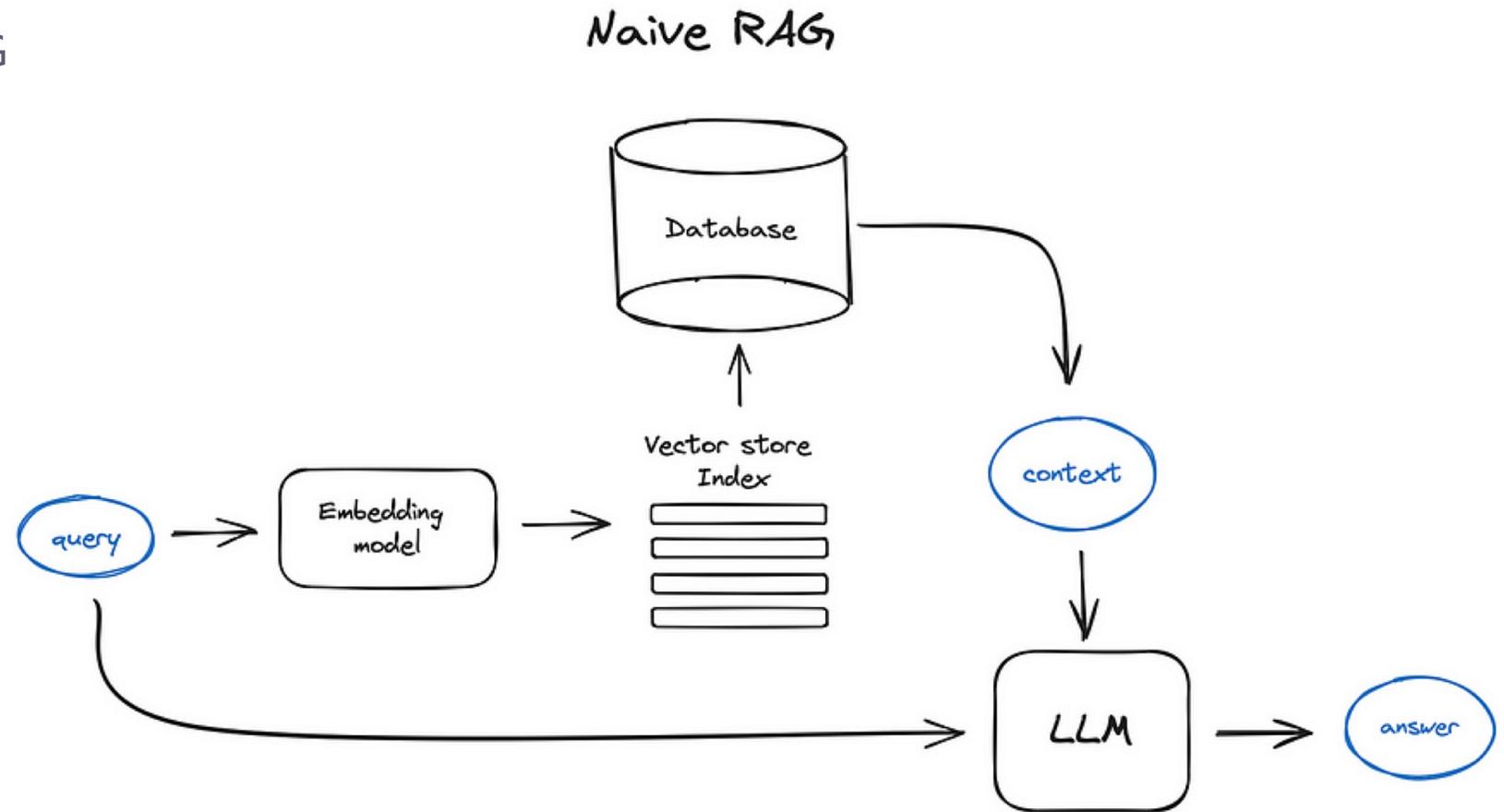
Зачем нужны embedding модели

- ▶ Задачи в ML делятся на генеративные и дискриминативные
- ▶ Baseline в большинстве дискриминативных задач – либо BERT, либо E5
- ▶ 72 млн скачиваний bert-base-uncased и 2 млн скачиваний multilingual-e5-large за последний месяц на hugging face



Retrieval Augmented Generation

Ключевая компонента RAG систем – это поисковой движок основанный на векторном поиске



```
{  
  "user_query": "How to use Microsoft Power BI for data analysis",  
  "positive_document": "Microsoft Power BI is a sophisticated tool that requires time and practice to master. In this tutorial, we'll show you how to navigate Power BI ... (omitted) ",  
  "hard_negative_document": "Excel is an incredibly powerful tool for managing and analyzing large amounts of data. Our tutorial series focuses on how you...(omitted)" }  
}
```

Задача модели выучить отображение так, чтобы вектор запроса был как можно ближе к искомому документу, чем к негативному



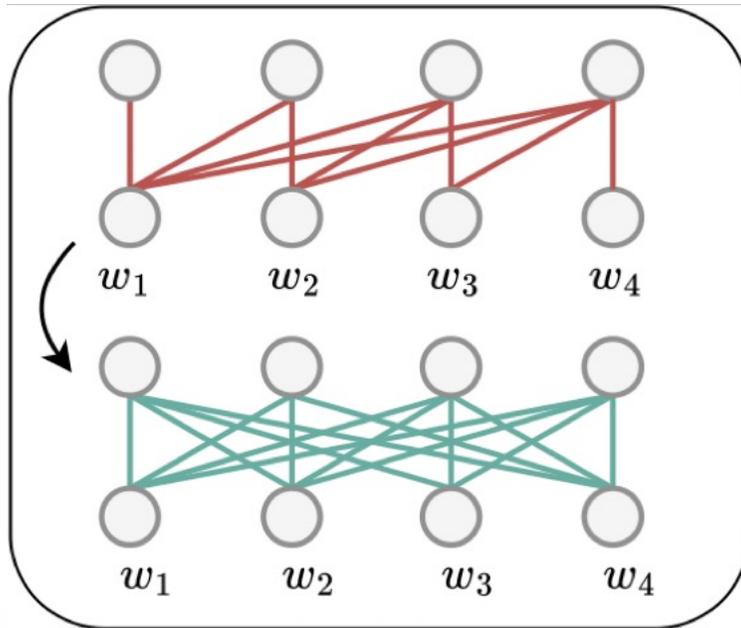
Архитектура

- ▶ Инициализация GigaChat-3B-pretrain с эмбедами RoPE, т.е. можно увеличивать размер контекста



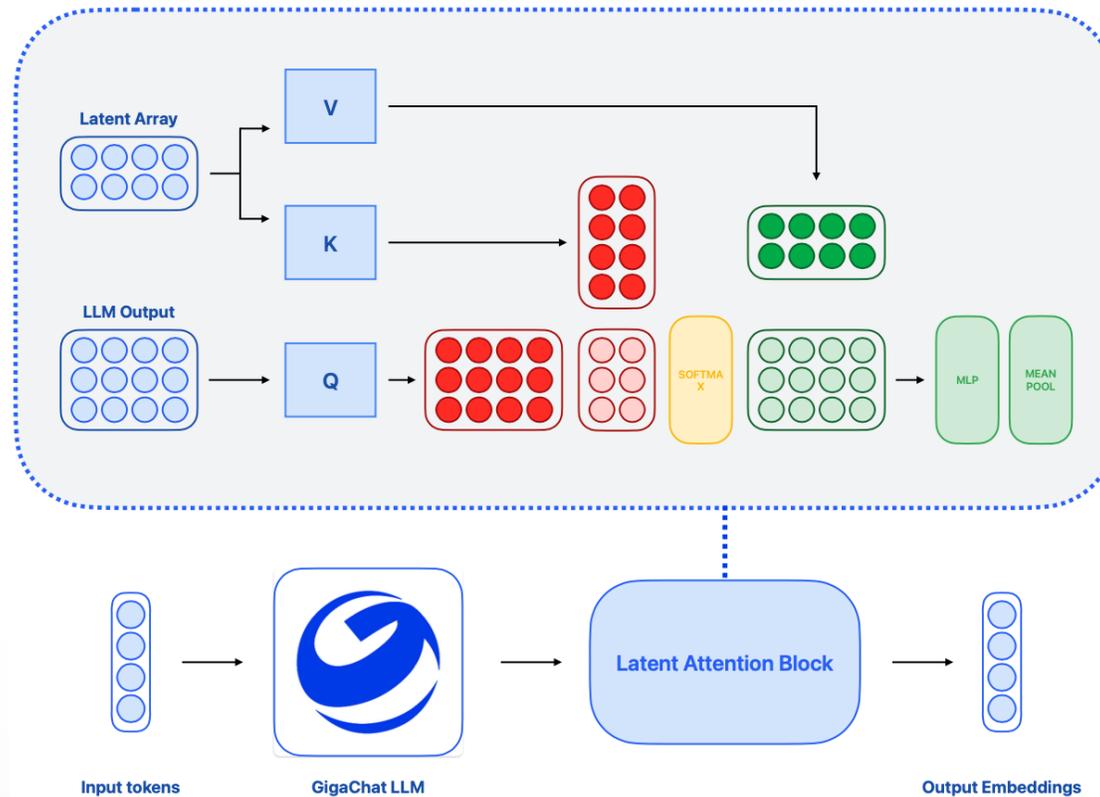
Архитектура

- ▶ Изменили механизм внимания LLM – с Decoder на Encoder



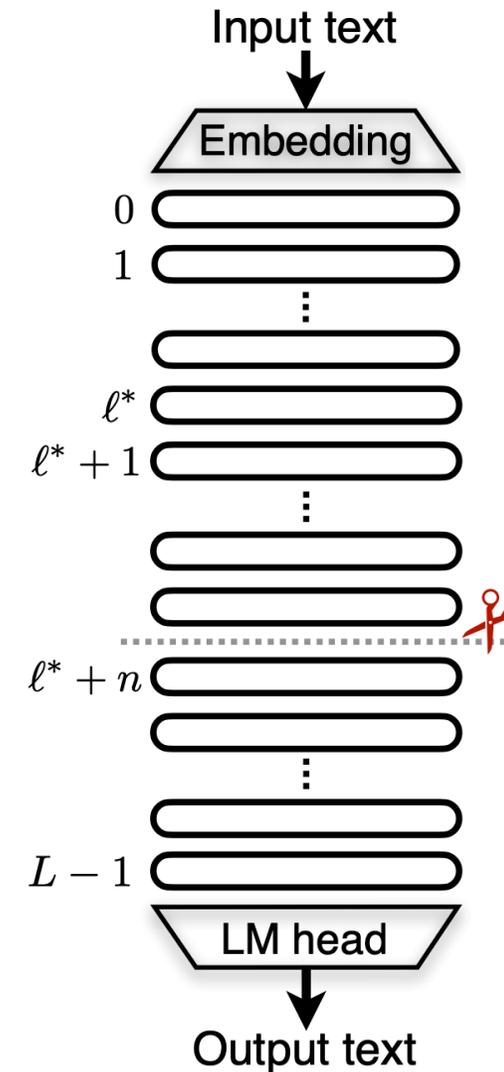
Архитектура

- ▶ Используется **Latent Attention Pooling** – как способ агрегации скрытых состояний для получения финального вектора



Архитектура

- ▶ Выбросили последние 9 из 36 трансформер блоков, что составляет 25% параметров исходной модели, практически без потери качества.



Пайплайн обучения модели

- ▶ **Pretrain** – обучаемся на большом кол-ве сырых retrieval данных
- ▶ **Fine-tuning** – дообучаемся на хороших чистых retrieval данных
- ▶ **Multitask** – добавляем другие задачи (классификация, sts, etc.) в обучение



	Pretrain	Fine-tune	Multitask
Data	Wiki, reddit, stackexchange, etc.	MS-MARCO, MIRACL, NQ, SQuAD, etc.	MS-MARCO, MIRACL, NQ, SQuAD + NLI, clf
Learning rate	3e-5	3e-5	2e-5
Batch size	16384	512	512
Hard negatives (per sample)	0	7	7
In-batch negatives	+	+	-



$$\min \mathbb{L} = -\log \frac{\phi(q_{\text{inst}}^+, d^+)}{\phi(q_{\text{inst}}^+, d^+) + \sum_{n_i \in \mathbb{N}} (\phi(q_{\text{inst}}^+, n_i))}$$

$$\phi(q, d) = \exp\left(\frac{1}{\tau} \cos(\mathbf{h}_q, \mathbf{h}_d)\right)$$

Для обучения используем классический InfoNCE loss



Инженерные трюки

- ▶ Fuse operations (RMSNorm, RoPE, flash-attn) – объединяем несколько операций в одну
- ▶ Эффективный чекпоинтинг активаций – сокращаем потребление GPU памяти
- ▶ Использовали алгоритм **gradient cache** для увеличения размера батча
- ▶ Использовали предобученные модели для поиска hard-негативов
- ▶ Эффективно переиспользуем pretrain модели – благодаря хорошей инициализации тратим несколько часов для обучения



Использование инструкций на естественном языке
позволяет повысить качество модели

как найти градиент функции



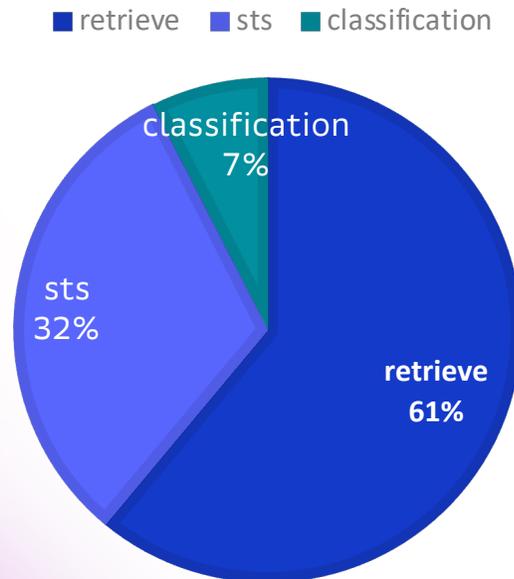
Дан вопрос, необходимо найти
абзац текста с ответом:
как найти градиент функции



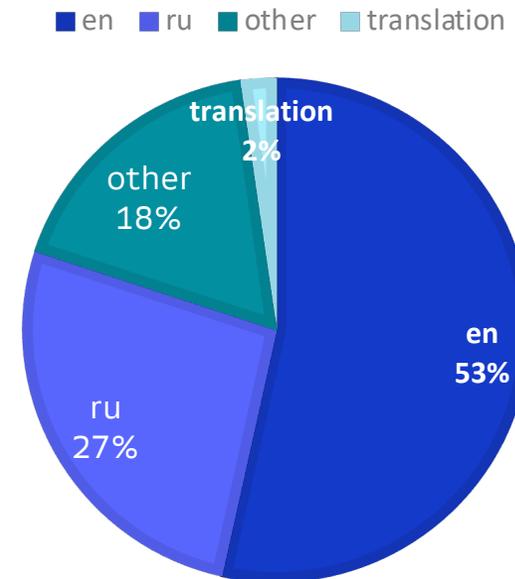
Данные

- ▶ Более **8** миллионов триплетов
- ▶ Более **50** различных источников данных были использованы для обучения модели

РАСПРЕДЕЛЕНИЕ ДАННЫХ ПО ТИПАМ ЗАДАЧ



РАСПРЕДЕЛЕНИЕ ДАННЫХ ПО ЯЗЫКАМ

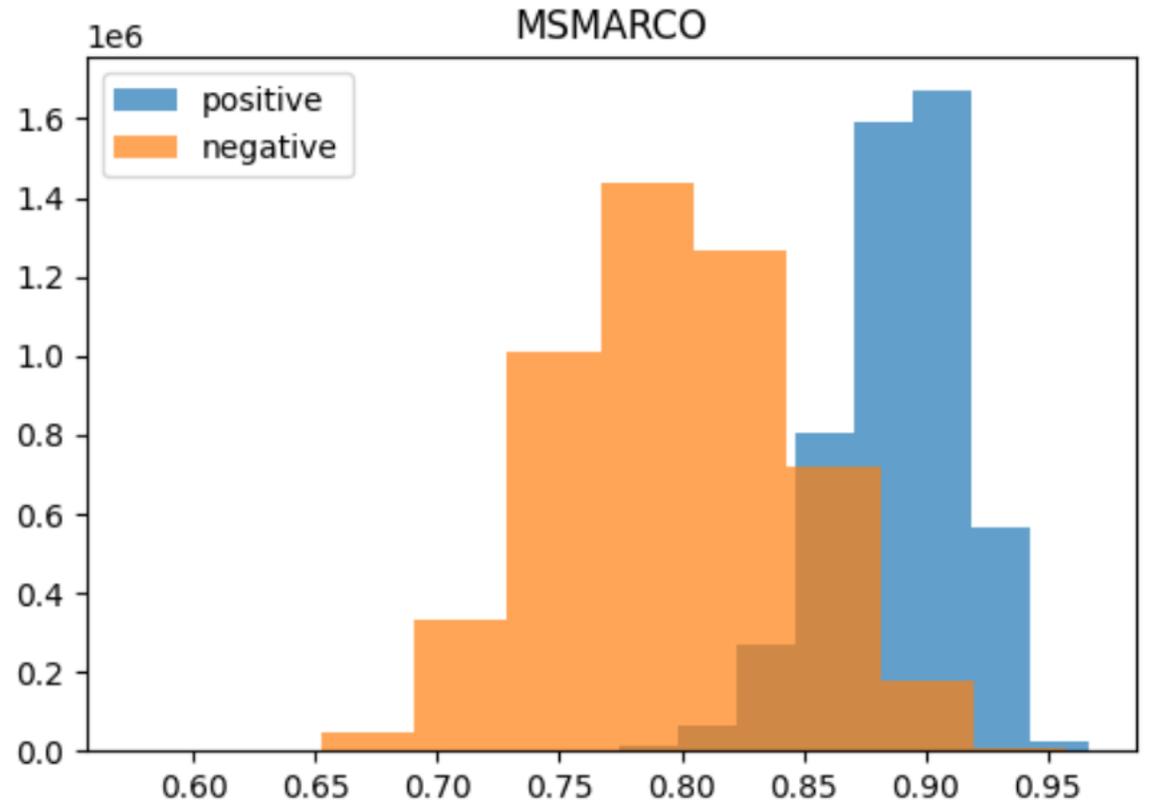


▶ Чистка открытых сетей с помощью e5-mistral-7b-instruct

Input: is a little caffeine ok during pregnancy

Positive: We don't know a lot about the effects of caffeine during pregnancy on you and your baby. So it's best to limit the amount you get each day. If you are pregnant, **limit caffeine to 200 milligrams each day**. This is about the amount in \hat{A} ½ 8-ounce cups of coffee or one 12-ounce cup of coffee.

Negative: It is generally safe for pregnant women to eat chocolate because studies have shown to prove certain benefits of eating chocolate during pregnancy. However, pregnant women should ensure their **caffeine intake is below 200 mg per day**.



- ▶ Переводы сетов на русский язык с помощью генераций GigaChat Max:
 - ▶ Только для задач retrieval
 - ▶ Только для текстов, не содержащих формулы, таблицы, код
 - ▶ Валидация переводов

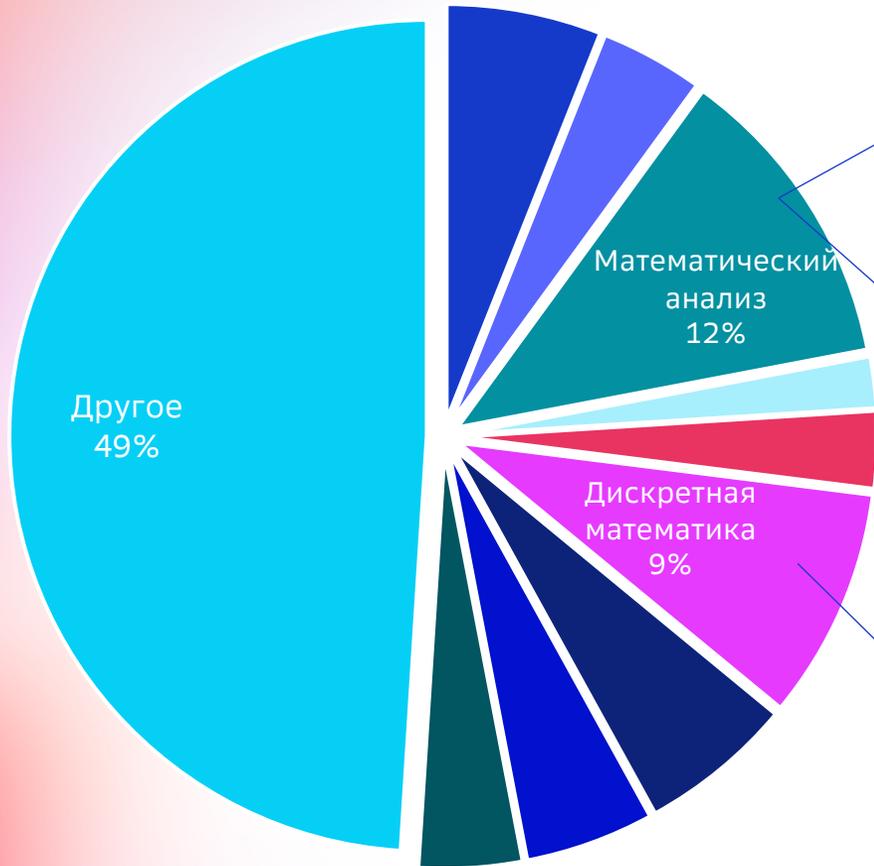


Открытые сети

- ▶ Расширение сетей с помощью генераций GigaChat 2 Max:
 - ▶ Только для задач retrieval
 - ▶ Переформулирование вопросов
 - ▶ Суммаризация позитивов и негативов
 - ▶ Переписывание позитивов и негативов



Кластеризация претрейна Математика



- Теория вероятностей и математическая статистика
- Теория оптимизации
- Математический анализ
- Численные методы
- Комплексный анализ
- Дискретная математика
- Линейная алгебра

Input: Производная функции – понятие дифференциального исчисления, характеризующее скорость изменения функции в данной точке. Определяется как предел отношения приращения функции к приращению её аргумента при стремлении приращения аргумента к нулю (при условии, что такой предел существует).

Positive: Дифференцируемая (в точке) функция – это функция, у которой существует дифференциал (в данной точке). Дифференцируемая на некотором множестве функция – это функция, дифференцируемая в каждой точке данного множества. Дифференцируемость является одним из фундаментальных понятий в математике.

Negative: Алгебра логики в её современном изложении занимается исследованием операций с высказываниями, то есть с предложениями, которые характеризуются только одним качеством – истинностным значением (истина, ложь). В классической алгебре логики высказывание одновременно может иметь только одно из двух истинностных значений: «истина» или «ложь».



Разметка

- ▶ Узкоспециализированные домены
- ▶ Использование RAG-функции на основе GigaEmbedder

В чем заключается паритет реальных процентных ставок?

Некорректный Требуется уточнение

1 rag

Функция для поиска релевантных пассажей.
Возвращает релевантные вопросу отрывки документов.
Дай максимально развернутый полный ответ на запрос пользователя, основываясь на релевантной информации из пассажей.
Если результат поиска не дал полезной информации, то отвечай на основе своих знаний или переспроси, чтобы уточнить.

^ Параметры

Поисковый запрос для ответа на вопрос.

***query:**
" В чем заключается паритет реальных процентных ставок?"

^ Результат

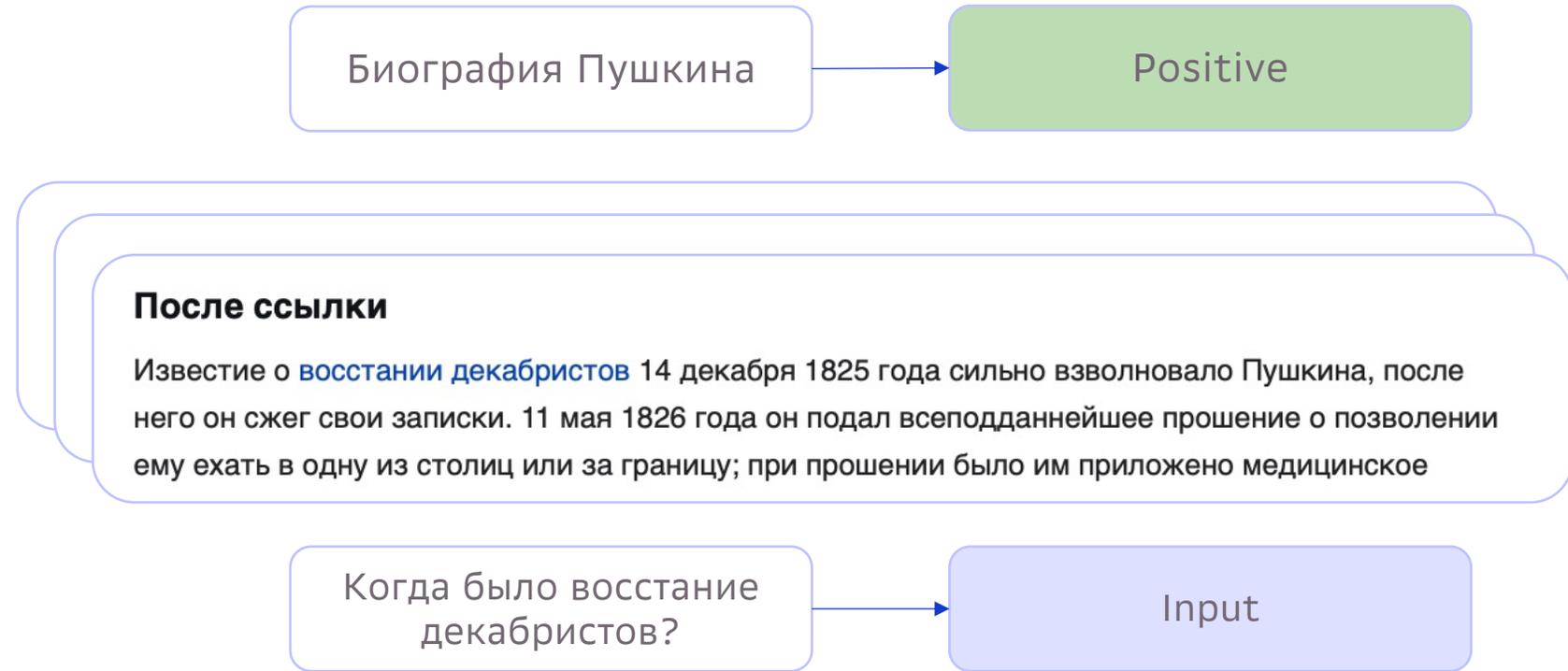
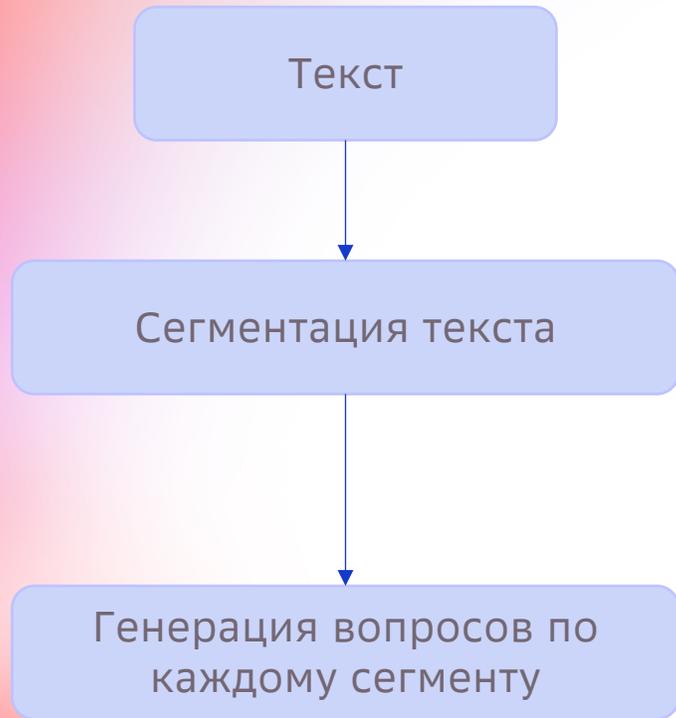
1:
" Паритет реальных процентных ставок В экономической теории проводится важное различие между номинальными процентными ставками, являющимися нормами дохода, измеренного деньгами, и реальными процентными ставками, являющимися нормами дохода, измеренного в реальном выражении, т.е. объемом выпуска продукции данной страны. Поскольку реальные процентные ставки часто неопределены, мы обычно говорим об ожидаемых реальных процентных ставках. Процентные ставки, которые мы обсуждали в связи с условием паритета процентных ставок и детерминантами спроса на деньги, были номинальными, например, долларовая доходность по долларовым депозитам. Но для многих других целей экономистам требуется анализировать поведение с позиций реального дохода. Ни одно лицо, намеревающееся инвестировать деньги, например, не могло бы принять решение, зная лишь, что номинальная процентная ставка равна . Инвестиция могла бы оказаться вполне привлекательной при нулевой инфляции, но катастрофически непривлекательной при инфляции, галопирующей с темпом в год! Мы завершаем эту главу, показывая, что когда согласно условию паритета номинальных процентных ставок разности номинальных процентных ставок по разным валютам приравниваются к ожидаемым изменениям номинальных валютных курсов, условие паритета реальных процентных ставок приравнивает разности ожидаемых реальных процентных ставок к ожидаемым изменениям реальных валютных курсов. Лишь когда ожидается, что относительный ППС будет соблюдаться (не ожидается изменения реального валютного курса), ожидаемые реальные процентные ставки во всех странах будут одинаковы. Ожидаемая реальная процентная ставка, обозначаемая , определяется как номинальная процентная ставка за вычетом ожидаемого темпа инфляции, : $r^{\{e\}} = R - \psi_1^{\{e\}}$."

2:
" $R = P + i + P \times i / 100$ Международный эффект Фишера. Выражает равновесие между существующими и будущими курсами спот и разницей процентных ставок в двух странах. Корректировка обменного курса будет эквивалентна и противоположна разнице процентных





Сбор хард-позитивов



После ссылки
Известие о [восстании декабристов](#) 14 декабря 1825 года сильно взволновало Пушкина, после него он сжег свои записки. 11 мая 1826 года он подал всеподданнейшее прошение о позволении ему ехать в одну из столиц или за границу; при прошении было им приложено медицинское



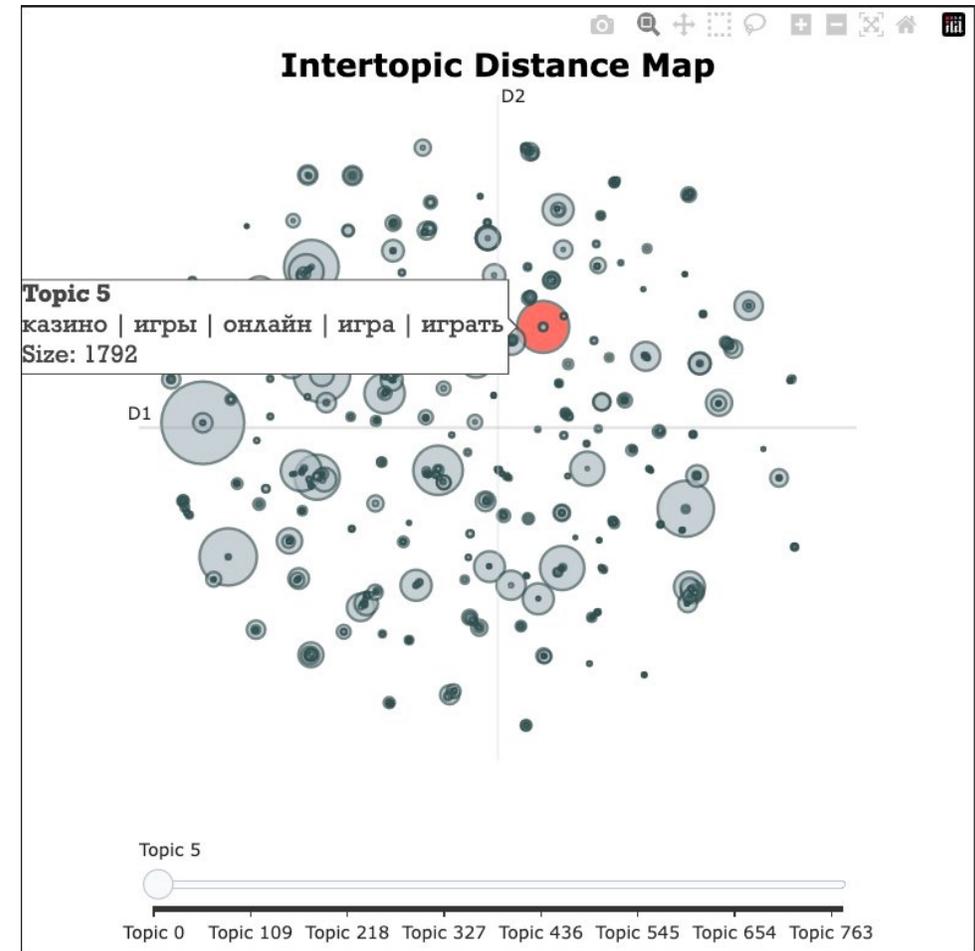
Результаты и метрики

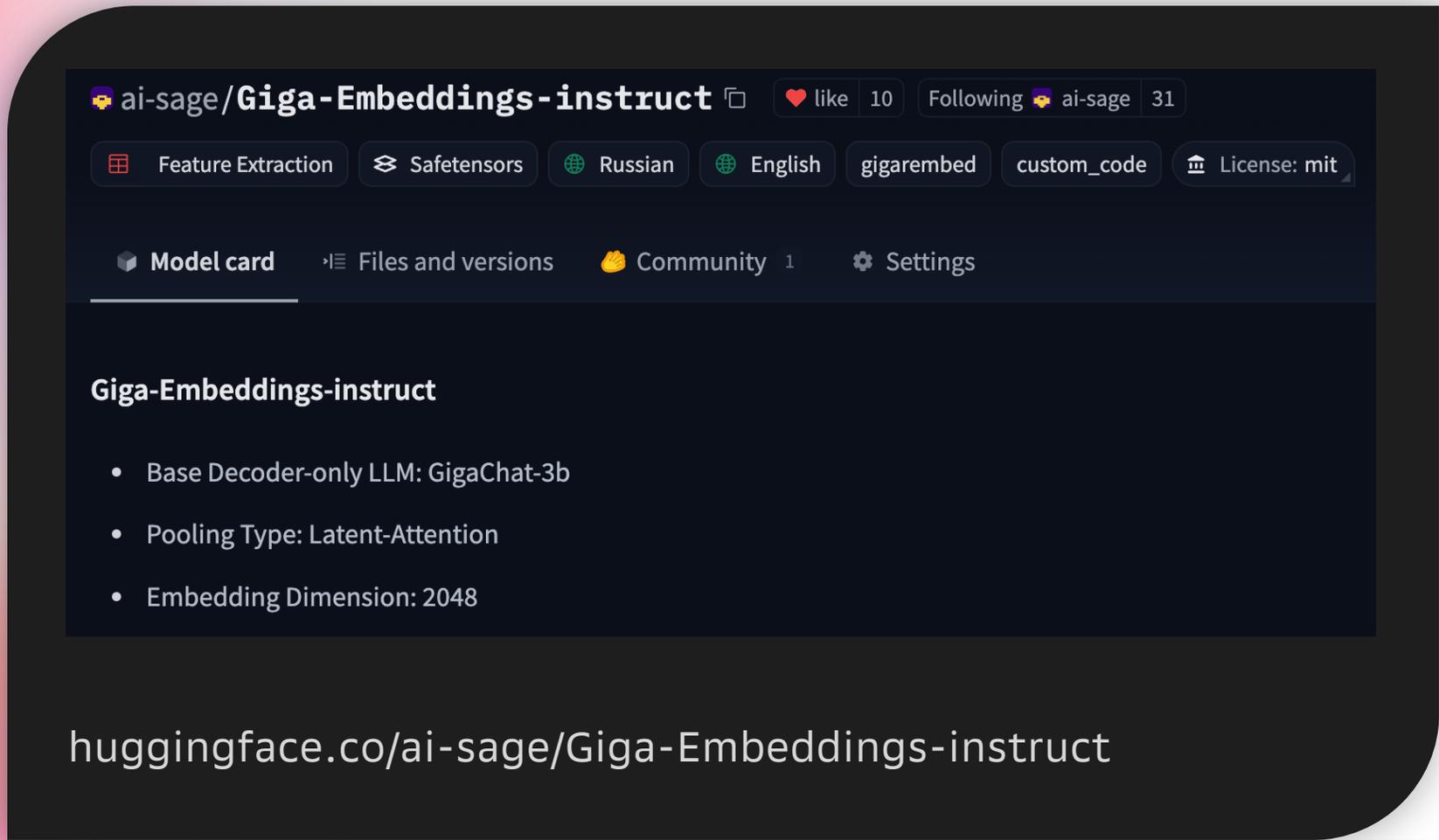
- ▶ Размер контекста **4096**
- ▶ Количество параметров **2B**

RuMTEB								
	Mean	Classification	Clustering	Multilabel classification	Pair classification	Reranking	Retrieval	STS
FRIDA	71.0	73.7	67.5	52.2	66.4	71.5	77.17	74.3
giga-embeddings-instruct	69.5	73.5	64.5	54.7	60.6	69.9	71.4	72.6
e5-mistral-7b-instruct	67.1	69.0	64.2	42.9	60.8	69.9	74.2	73.7
multilingual-e5-large-instruct	65.0	66.3	63.1	41.1	63.9	64.3	68.2	76.5



- ▶ Бизнес-задачи с использованием RAG
 - ▶ Работа с документами
 - ▶ Обработка новостей
- ▶ Текстовый энкодер для модели генерации изображений Kandinsky
- ▶ Ранжировщик функций для Ассистента на GigaChat
- ▶ Аналитические задачи
 - ▶ Кластеризация
 - ▶ Дедупликация





The screenshot shows the Hugging Face interface for the model `ai-sage/Giga-Embeddings-instruct`. At the top, it displays the model name, a copy icon, and engagement metrics: 10 likes and 31 followers. Below this are several filter tags: Feature Extraction, Safetensors, Russian, English, gigarembed, custom_code, and License: mit. The navigation bar includes Model card (selected), Files and versions, Community (1), and Settings. The main content area is titled "Giga-Embeddings-instruct" and lists the following specifications:

- Base Decoder-only LLM: GigaChat-3b
- Pooling Type: Latent-Attention
- Embedding Dimension: 2048

At the bottom of the screenshot, the URL `huggingface.co/ai-sage/Giga-Embeddings-instruct` is displayed.



**GETGGA
CHAT**