

точка банк



ТАЧКА НА ПРОКАЧКУ

Адаптируем вашу старую ретриву к новым вызовам





АГЕНДА

1. Как выглядит стандартная ретрива?
2. Зачем что-то менять?
3. Как улучшить качество модели?
4. Как ускорить систему?
5. Что делать с ростом количества источников?

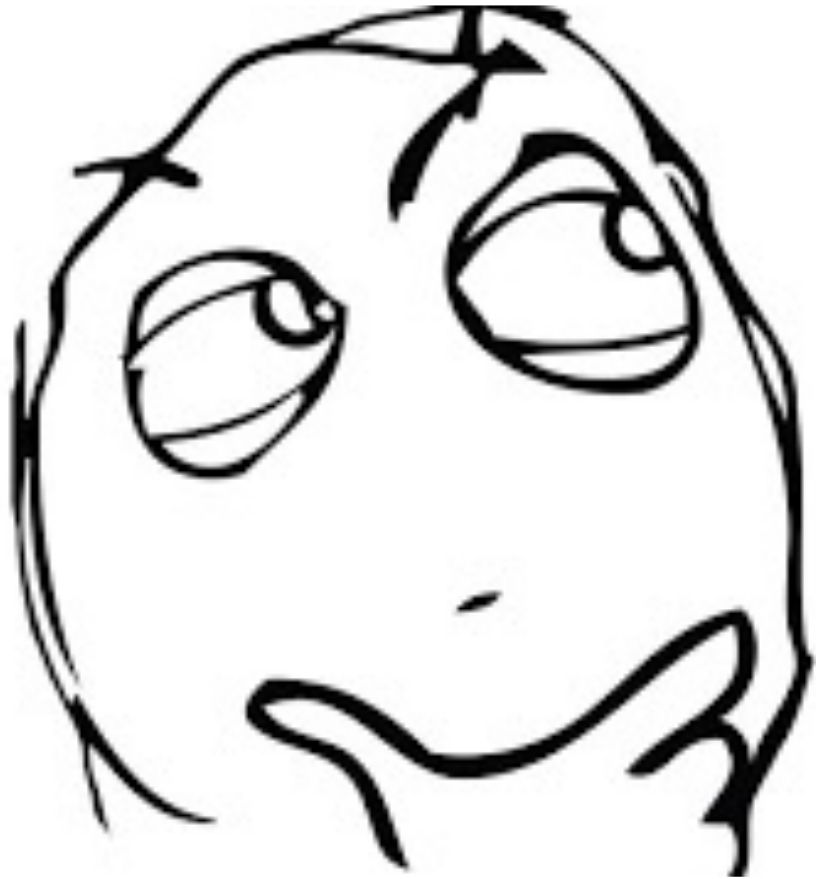
ВЛАД ПОПОВ

TechLead Foundation Models



ЧТО ТАКОЕ РЕТРИВА?

точка банк



ЧТО ТАКОЕ РЕТРИВА?

точка банк

Фильм про человека с задержкой в развитии, который бегал и стал миллионером.

ЧТО ТАКОЕ РЕТРИВА?

точка банк

Фильм про человека с задержкой в развитии, который бегал и стал миллионером.

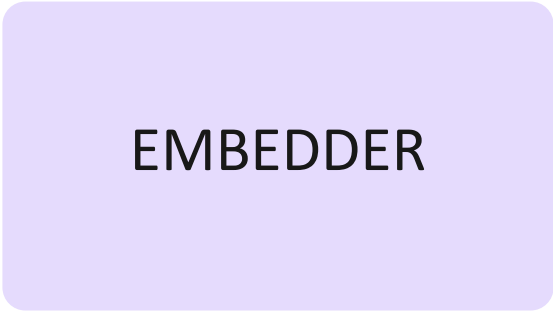


КАК ВЫГЛЯДИТ СТАНДАРТНАЯ РЕТРИВА В 2026 ГОДУ

СТАНДАРТНАЯ РЕТРИВА

точка банк

Компоненты



EMBEDDER

СТАНДАРТНАЯ РЕТРИВА

точка банк

Пример выдачи: NN embedder

как ИП на УСН уменьшить налог на страховые взносы



Предприниматель на упрощёнке без работников может целиком зачесть в счёт платежа в бюджет деньги, отправленные в Соцфонд за себя. Если отчислений хватает, платить в казну не придётся ничего.

СТАНДАРТНАЯ РЕТРИВА

точка банк

Компоненты

EMBEDDER

+

TF-IDF
BM25
SPLADE

СТАНДАРТНАЯ РЕТРИВА

точка банк

Пример выдачи: statistical embedder

как ИП на УСН уменьшить налог на страховые взносы



Страховые взносы ИП на УСН в 2026 году: фиксированные взносы ИП, сроки уплаты взносов и КБК. ИП на УСН платит взносы даже при нулевом доходе.

СТАНДАРТНАЯ РЕТРИВА

точка банк

Компоненты



СТАНДАРТНАЯ РЕТРИВА

точка банк

Пример выдачи: NN reranker

как ИП на УСН уменьшить налог на страховые взносы



Предприниматель на упрощёнке без работников может целиком зачесть в счёт платежа в бюджет деньги, отправленные в Соцфонд за себя. Если отчислений хватает, платить в казну не придётся ничего.

**ЗАЧЕМ ЧТО-ТО
МЕНЯТЬ?**

ЗАЧЕМ ЧТО-ТО МЕНЯТЬ?

точка банк

Большему количеству сервисов нужен «поисковик»

ЗАЧЕМ ЧТО-ТО МЕНЯТЬ?

точка банк

Большому количеству сервисов нужен «поисковик»

- Рост требований к качеству

ЗАЧЕМ ЧТО-ТО МЕНЯТЬ?

точка банк

Большому количеству сервисов нужен «поисковик»

- Рост требований к качеству
- Рост требований к производительности

ЗАЧЕМ ЧТО-ТО МЕНЯТЬ?

точка банк

Большому количеству сервисов нужен «поисковик»

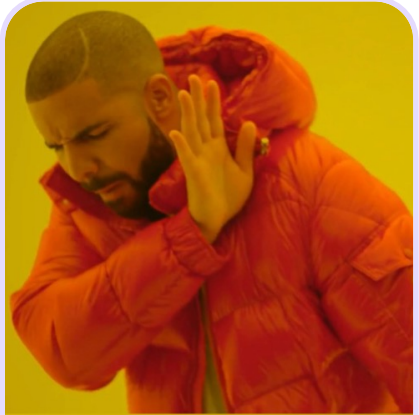
- Рост требований к качеству
- Рост требований к производительности
- Рост количества источников

КАК ПРОКАЧАТЬ КАЧЕСТВО РЕТРИВЫ

КАК ПРОКАЧАТЬ КАЧЕСТВО РЕТРИВЫ

точка банк

Как оценить качество?



Кросс-
валидация



Бенчмарки

КАК ПРОКАЧАТЬ КАЧЕСТВО РЕТРИВЫ

точка банк

Бенчмарки

MTEB
Massive Text
Embedding Benchmark



Beir
Benchmarking IR

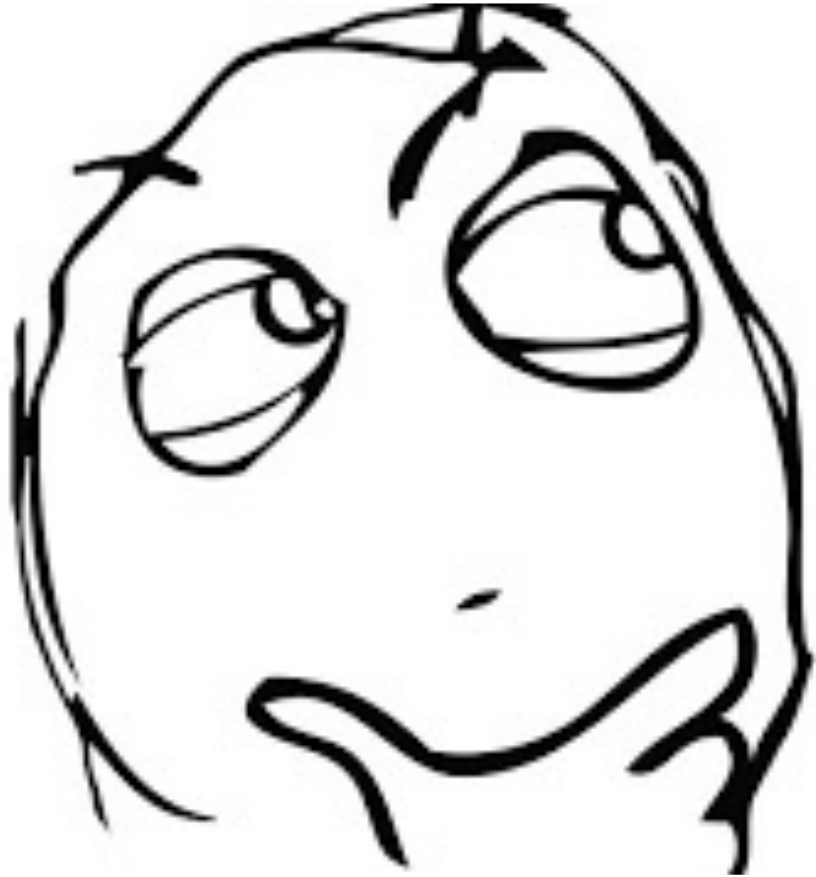
MS MARCO

Natural Questions

КАК ПРОКАЧАТЬ КАЧЕСТВО РЕТРИВЫ

точка банк

Свой бенчмарк:



КАК ПРОКАЧАТЬ КАЧЕСТВО РЕТРИВЫ

точка банк

Свой бенчмарк:

1. История запросов

КАК ПРОКАЧАТЬ КАЧЕСТВО РЕТРИВЫ

точка банк

Свой бенчмарк:

1. История запросов
2. Синтетические данные

КАК ПРОКАЧАТЬ КАЧЕСТВО РЕТРИВЫ

точка банк

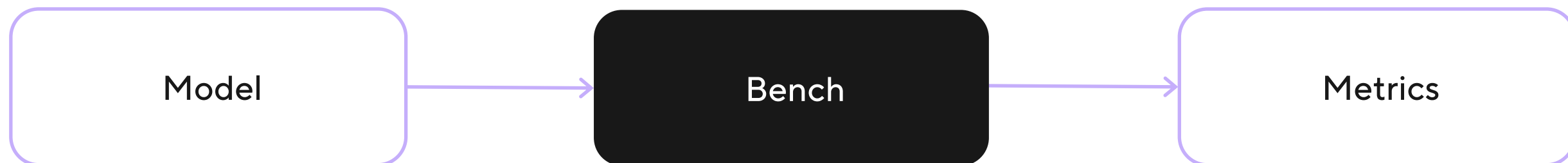
Свой бенчмарк:

1. История запросов
2. Синтетические данные
3. Конкатенация из других бенчей

КАК ПРОКАЧАТЬ КАЧЕСТВО РЕТРИВЫ

точка банк

Свой бенчмарк:



КАК ПРОКАЧАТЬ КАЧЕСТВО РЕТРИВЫ

точка банк

Свой бенчмарк:

1. recall@k
2. hitrate@k
3. mrr@k

КАК ПРОКАЧАТЬ КАЧЕСТВО РЕТРИВЫ

точка банк

Свой бенчмарк:

Macro-metrics

Отслеживать общее качество
Легко интерпретировать

Micro-metrics

Больше деталей
Больше возможностей
увидеть пробелы
Сложнее интерпретировать

КАК ПРОКАЧАТЬ КАЧЕСТВО РЕТРИВЫ

точка банк

Свой бенчмарк: predprinimatel.ru

Macro-metrics

hitrate@1=71%

Micro-metrics

hitrate_questions@1=0.72

hitrate_keywords@1=0.71

hitrate_topic@1=0.74

hitrate_quote@1=0.71

КАК ПРОКАЧАТЬ КАЧЕСТВО РЕТРИВЫ

точка банк

Как НЕ улучшить качество?

1. Взять модель побольше

КАК ПРОКАЧАТЬ КАЧЕСТВО РЕТРИВЫ

точка банк

Как НЕ улучшить качество?

1. Взять модель побольше — на вашем домене может не сработать + много ресурсов

КАК ПРОКАЧАТЬ КАЧЕСТВО РЕТРИВЫ

точка банк

Как НЕ улучшить качество?

1. Взять модель побольше — на вашем домене может не сработать + много ресурсов
2. Увеличить top-k

КАК ПРОКАЧАТЬ КАЧЕСТВО РЕТРИВЫ

точка банк

Как НЕ улучшить качество?

1. Взять модель побольше — на вашем домене может не сработать + много ресурсов
2. Увеличить top-k — релеватный документ может так и не появиться

КАК ПРОКАЧАТЬ КАЧЕСТВО РЕТРИВЫ

точка банк

Как НЕ улучшить качество?

1. Взять модель побольше — на вашем домене может не сработать + много ресурсов
2. Увеличить top-k — релеватный документ может так и не появиться
3. Обучение на general data

КАК ПРОКАЧАТЬ КАЧЕСТВО РЕТРИВЫ

точка банк

Как НЕ улучшить качество?

1. Взять модель побольше — на вашем домене может не сработать + много ресурсов
2. Увеличить top-k — релеватный документ может так и не появиться
3. Обучение на general data — модель не выучится различать тонкости

КАК ПРОКАЧАТЬ КАЧЕСТВО РЕТРИВЫ

точка банк

Как улучшить качество?



Энкодеры
~миллиарды токенов



Декодеры
~триллионы токенов

КАК ПРОКАЧАТЬ КАЧЕСТВО РЕТРИВЫ

точка банк

Как улучшить качество?



Энкодеры
~миллиарды токенов

Декодеры
~триллионы токенов

КАК ПРОКАЧАТЬ КАЧЕСТВО РЕТРИВЫ

точка банк

Что получаем от конверсии LLM в embedder?

КАК ПРОКАЧАТЬ КАЧЕСТВО РЕТРИВЫ

точка банк

Что получаем от конверсии LLM в embedder?

1. Лучше понимаем «человеческие» запросы

КАК ПРОКАЧАТЬ КАЧЕСТВО РЕТРИВЫ

точка банк

Что получаем от конверсии LLM в embedder?

1. Лучше понимаем «человеческие» запросы
2. Умеем в инструкции

КАК ПРОКАЧАТЬ КАЧЕСТВО РЕТРИВЫ

точка банк

Что получаем от конверсии LLM в embedder?

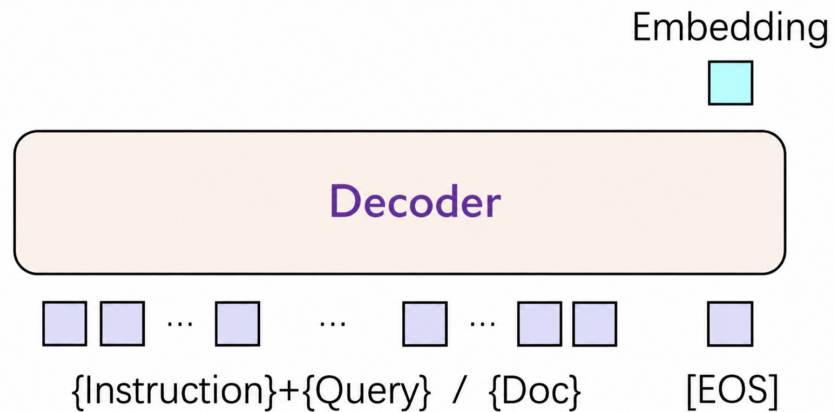
1. Лучше понимаем «человеческие» запросы
2. Умеем в инструкции
3. Поддерживаем оптимизации в виде KV-caching, flash attention, etc...

КАК ПРОКАЧАТЬ КАЧЕСТВО РЕТРИВЫ

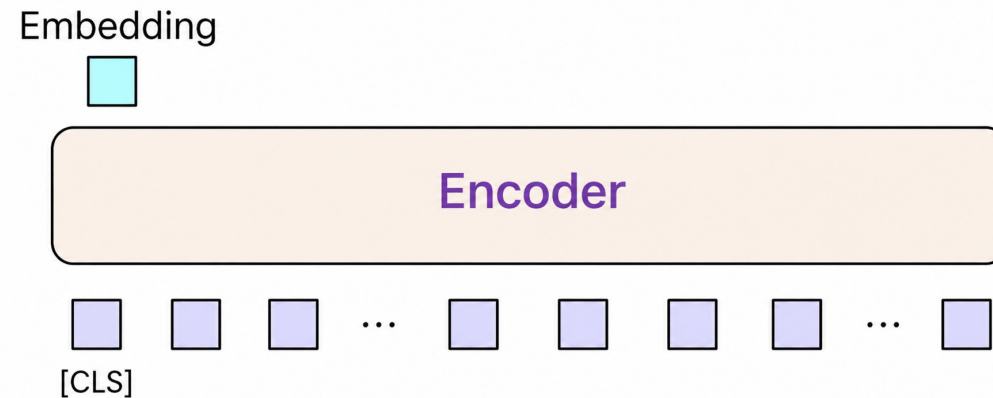
точка банк

Как улучшить качество?

Декодер



Энкодер



КАК ПРОКАЧАТЬ КАЧЕСТВО РЕТРИВЫ

точка банк

Где найти данные для обучения?

1. Логи
2. Синтетические данные
3. Интернет (сайты, форумы)

КАК ПРОКАЧАТЬ КАЧЕСТВО РЕТРИВЫ

точка банк

Как собрать датасет?

КАК ПРОКАЧАТЬ КАЧЕСТВО РЕТРИВЫ

точка банк

Как собрать датасет?

1. Найти/создать запросы к документа

КАК ПРОКАЧАТЬ КАЧЕСТВО РЕТРИВЫ

точка банк

Как собрать датасет?

1. Найти/создать запросы к документа
2. Аугментация запросов

КАК ПРОКАЧАТЬ КАЧЕСТВО РЕТРИВЫ

точка банк

Как собрать датасет?

1. Найти/создать запросы к документа
2. Аугментация запросов
3. Проверка корректности созданных связей

КАК ПРОКАЧАТЬ КАЧЕСТВО РЕТРИВЫ

точка банк

Как собрать датасет?

1. Найти/создать запросы к документа
2. Аугментация запросов
3. Проверка корректности созданных связей
4. Проверка на утечки из бенча

КАК ПРОКАЧАТЬ КАЧЕСТВО РЕТРИВЫ

точка банк

Как собрать датасет?

1. Найти/создать запросы к документа
2. Аугментация запросов
3. Проверка корректности созданных связей
4. Проверка на утечки из бенча
5. Собрать Hard-Negatives

КАК ПРОКАЧАТЬ КАЧЕСТВО РЕТРИВЫ

точка банк

Что такое Hard-Negative:

как ИП на УСН уменьшить налог на страховые взносы

```
graph TD; A[как ИП на УСН уменьшить налог на страховые взносы] --> B[ИП на УСН «доходы» уменьшают сумму налога на уплаченные страховые взносы. ИП без работников могут уменьшить налог на всю сумму взносов.]; A --> C[ИП на УСН «доходы минус расходы» не уменьшают налог на страховые взносы: взносы учитываются в составе расходов и уменьшают налоговую базу.];
```

ИП на УСН «доходы» уменьшают сумму налога на уплаченные страховые взносы. ИП без работников могут уменьшить налог на всю сумму взносов.

ИП на УСН «доходы минус расходы» не уменьшают налог на страховые взносы: взносы учитываются в составе расходов и уменьшают налоговую базу.

КАК ПРОКАЧАТЬ КАЧЕСТВО РЕТРИВЫ

точка банк

Hard-Negative mining:

1. Прогнать модель по корпусу

КАК ПРОКАЧАТЬ КАЧЕСТВО РЕТРИВЫ

точка банк

Hard-Negative mining:

1. Прогнать модель по корпусу
2. Для каждого запроса найти документы, которые ранжируются высоко

КАК ПРОКАЧАТЬ КАЧЕСТВО РЕТРИВЫ

точка банк

Hard-Negative mining:

1. Прогнать модель по корпусу
2. Для каждого запроса найти документы, которые ранжируются высоко
3. Документы, которые не пересекаются с разметкой, добавить как Hard-Negative

КАК ПРОКАЧАТЬ КАЧЕСТВО РЕТРИВЫ

точка банк

Итоги

Было:

- recall@1: 0.43
- recall@3: 0.58
- recall@10: 0.71

Стало:

- recall@1: 0.66
- recall@3: 0.69
- recall@10: 0.84
- hitrate@1: 0.66
- hitrate@3: 0.85
- hitrate@10: 0.95



КАК ПРОКАЧАТЬ ПРОИЗВОДИТЕЛЬНОСТЬ РЕТРИВЫ

КАК ПРОКАЧАТЬ СКОРОСТЬ РЕТРИВЫ

точка банк

Как мы НЕ будем делать:

1. Завалить компьютером

КАК ПРОКАЧАТЬ СКОРОСТЬ РЕТРИВЫ

точка банк

Как мы НЕ будем делать:

1. Завалить компьютером — у нас не бесконечное количество денег

КАК ПРОКАЧАТЬ СКОРОСТЬ РЕТРИВЫ

точка банк

Как мы НЕ будем делать:

1. Завалить компьютером — у нас не бесконечное количество денег
2. Выкидывать реранкер

КАК ПРОКАЧАТЬ СКОРОСТЬ РЕТРИВЫ

точка банк

Как мы НЕ будем делать:

1. Завалить компьютером — у нас не бесконечное количество денег
2. Выкидывать реранкер — не сможем фильтровать релевантные доки

КАК ПРОКАЧАТЬ СКОРОСТЬ РЕТРИВЫ

точка банк

Как мы НЕ будем делать:

1. Завалить компьютером — у нас не бесконечное количество денег
2. Выкидывать реранкер — не сможем фильтровать релевантные доки
3. Пережимать текущие модели

КАК ПРОКАЧАТЬ СКОРОСТЬ РЕТРИВЫ

точка банк

Как мы НЕ будем делать:

1. Завалить компьютером — у нас не бесконечное количество денег
2. Выкидывать реранкер — не сможем фильтровать релевантные доки
3. Пережимать текущие модели — качество сильно упадет

КАК ПРОКАЧАТЬ СКОРОСТЬ РЕТРИВЫ

точка банк

Среды для инференса эмбеддеров:



КАК ПРОКАЧАТЬ СКОРОСТЬ РЕТРИВЫ

точка банк

Среды для инференса:

microsoft / onnxruntime

 Code  Issues 826  Pull requests 530

sgl-project / sglang

 Code  Issues 597  Pull requests 2.1k

КАК ПРОКАЧАТЬ СКОРОСТЬ РЕТРИВЫ

точка банк

Среды для инференса энкодеров:



КАК ПРОКАЧАТЬ СКОРОСТЬ РЕТРИВЫ

точка банк

Среды для инференса декодеров:



КАК ПРОКАЧАТЬ СКОРОСТЬ РЕТРИВЫ

точка банк

Среды для инференса декодеров:

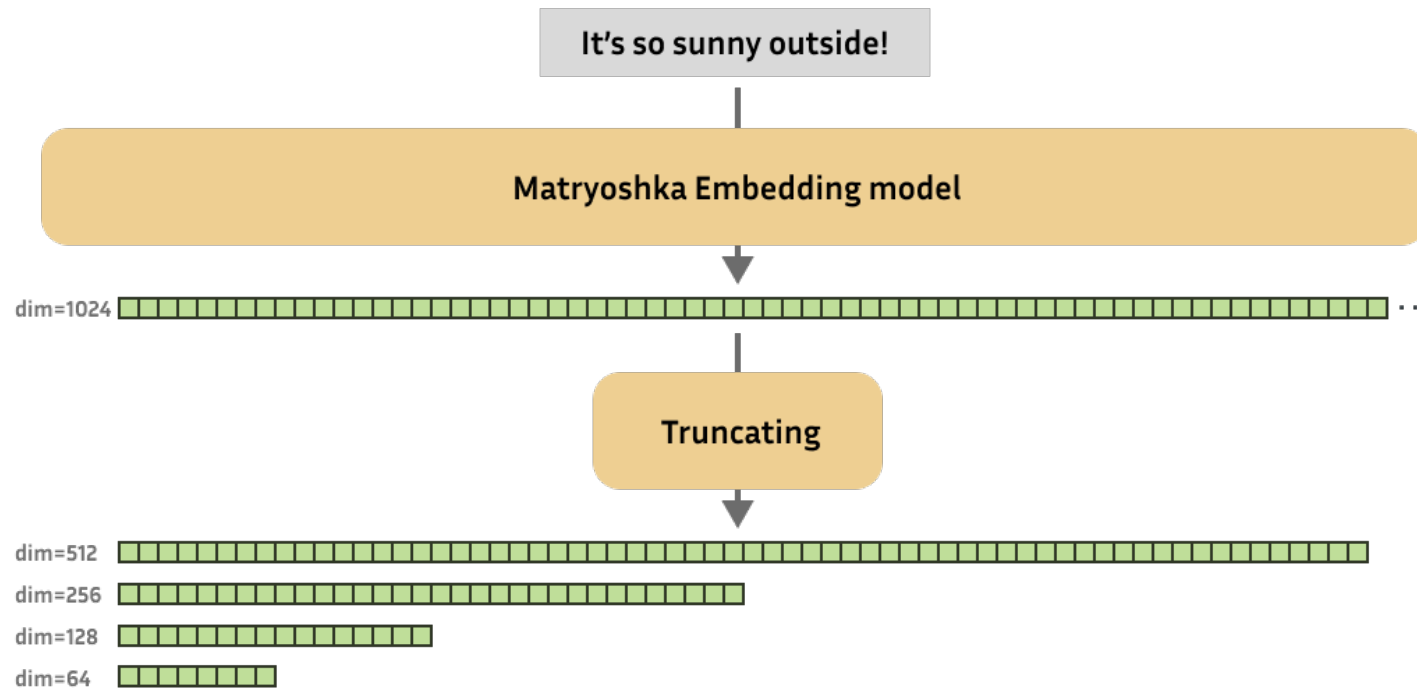
- Flash attention
- Prefix-caching

КАК ПРОКАЧАТЬ СКОРОСТЬ РЕТРИВЫ

точка банк

Что мы сейчас не применяем, но потенциально может быть полезно:

- Matryoshka Representation Learning



КАК ПРОКАЧАТЬ СКОРОСТЬ РЕТРИВЫ

точка банк

Что мы сейчас не применяем, но потенциально может быть полезно:

- Дистилляция огромных моделей



КАК ПРОКАЧАТЬ СКОРОСТЬ РЕТРИВЫ

точка банк

Итоги

Что у нас было:

- encoding latency: 5ms
- rerank depth: 5 docs

Стало:

- 1ms
- 20 docs



**ЧТО ДЕЛАТЬ С РОСТОМ
КОЛИЧЕСТВА
ИСТОЧНИКОВ?**

ЧТО ДЕЛАТЬ С РОСТОМ КОЛИЧЕСТВА ИСТОЧНИКОВ

точка банк

Как делать НЕ надо:

1. Использовать 100500 разных моделей на разные источники



ЧТО ДЕЛАТЬ С РОСТОМ КОЛИЧЕСТВА ИСТОЧНИКОВ

точка банк

Как делать НЕ надо:

1. Использовать 100500 разных моделей на разные источники
2. Игнорировать структуры документов



ЧТО ДЕЛАТЬ С РОСТОМ КОЛИЧЕСТВА ИСТОЧНИКОВ

точка банк

Как делать НЕ надо:

1. Использовать 100500 разных моделей на разные источники
2. Игнорировать структуры документов
3. Создавать универсальный индекс



ЧТО ДЕЛАТЬ С РОСТОМ КОЛИЧЕСТВА ИСТОЧНИКОВ

точка банк

Как делать НЕ надо:

1. Использовать 100500 разных моделей на разные источники
2. Игнорировать структуры документов
3. Создавать универсальный индекс
4. Игнорировать метаинформацию



ЧТО ДЕЛАТЬ С РОСТОМ КОЛИЧЕСТВА ИСТОЧНИКОВ

точка банк

Как делать НЕ надо:

1. Использовать 100500 разных моделей на разные источники
2. Игнорировать структуры документов
3. Создавать универсальный индекс
4. Игнорировать метаинформацию
5. Постоянно перестраивать индекс



ЧТО ДЕЛАТЬ С РОСТОМ КОЛИЧЕСТВА ИСТОЧНИКОВ

точка банк

Как делать надо:

1. Использовать единую general-модель

ЧТО ДЕЛАТЬ С РОСТОМ КОЛИЧЕСТВА ИСТОЧНИКОВ

точка банк

Как делать надо:

1. Использовать единую general-модель
2. Structure-aware chunking

ЧТО ДЕЛАТЬ С РОСТОМ КОЛИЧЕСТВА ИСТОЧНИКОВ

точка банк

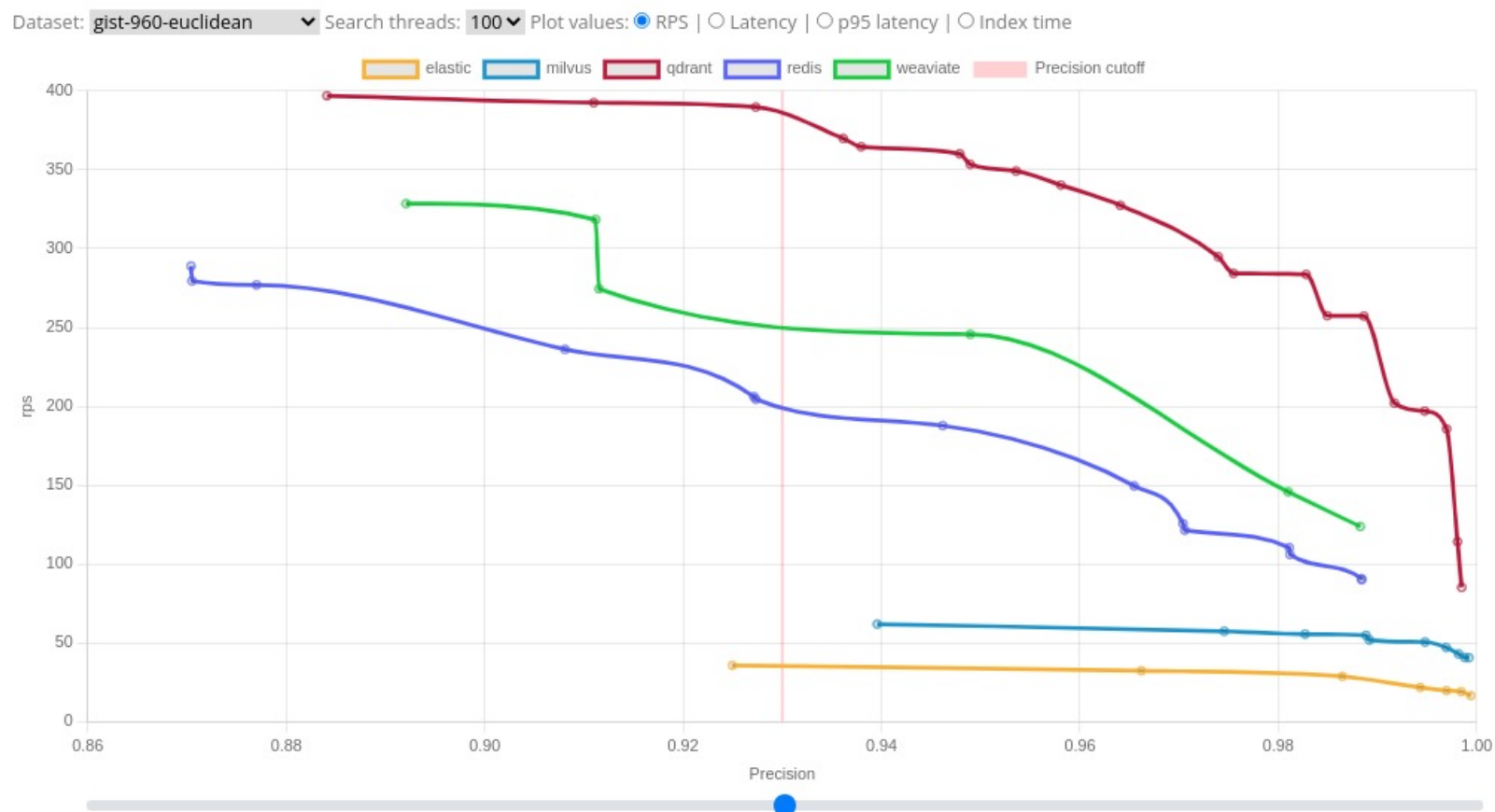
Как делать надо:

1. Использовать единую general-модель
2. Structure-aware chunking
3. Per-service index

ЧТО ДЕЛАТЬ С РОСТОМ КОЛИЧЕСТВА ИСТОЧНИКОВ

точка банк

VDB:

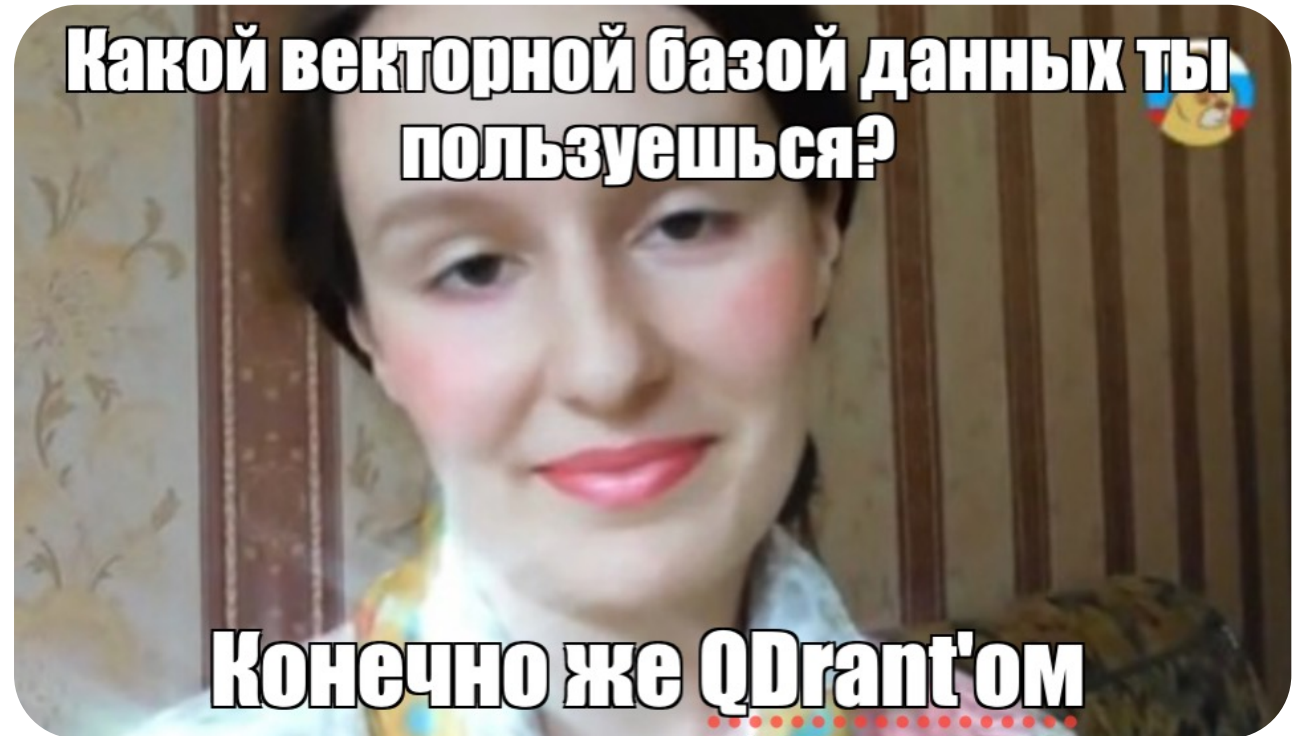


ЧТО ДЕЛАТЬ С РОСТОМ КОЛИЧЕСТВА ИСТОЧНИКОВ

точка банк

Как делать надо:

1. Использовать единую general-модель
2. Structure-aware chunking
3. Per-service index

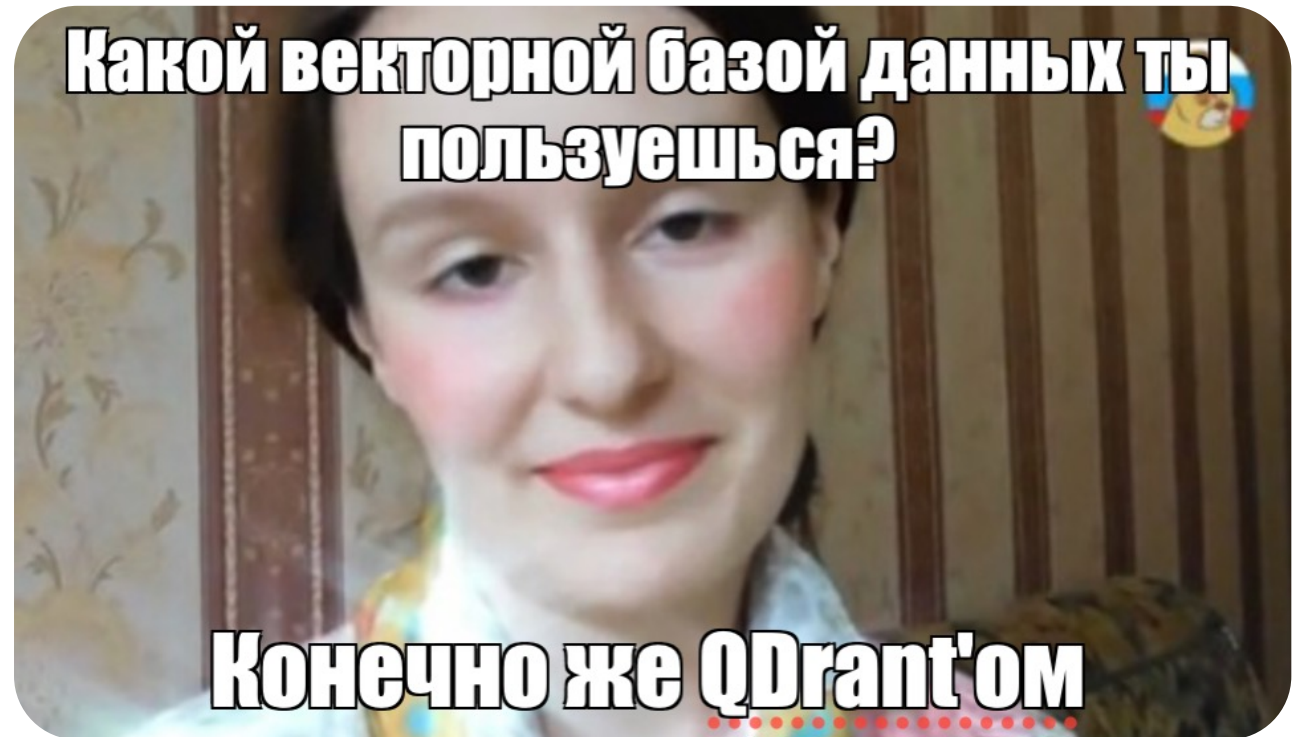


ЧТО ДЕЛАТЬ С РОСТОМ КОЛИЧЕСТВА ИСТОЧНИКОВ

точка банк

Как делать надо:

1. Использовать единую general-модель
2. Structure-aware chunking
3. Per-service index
4. Использовать метаданные!

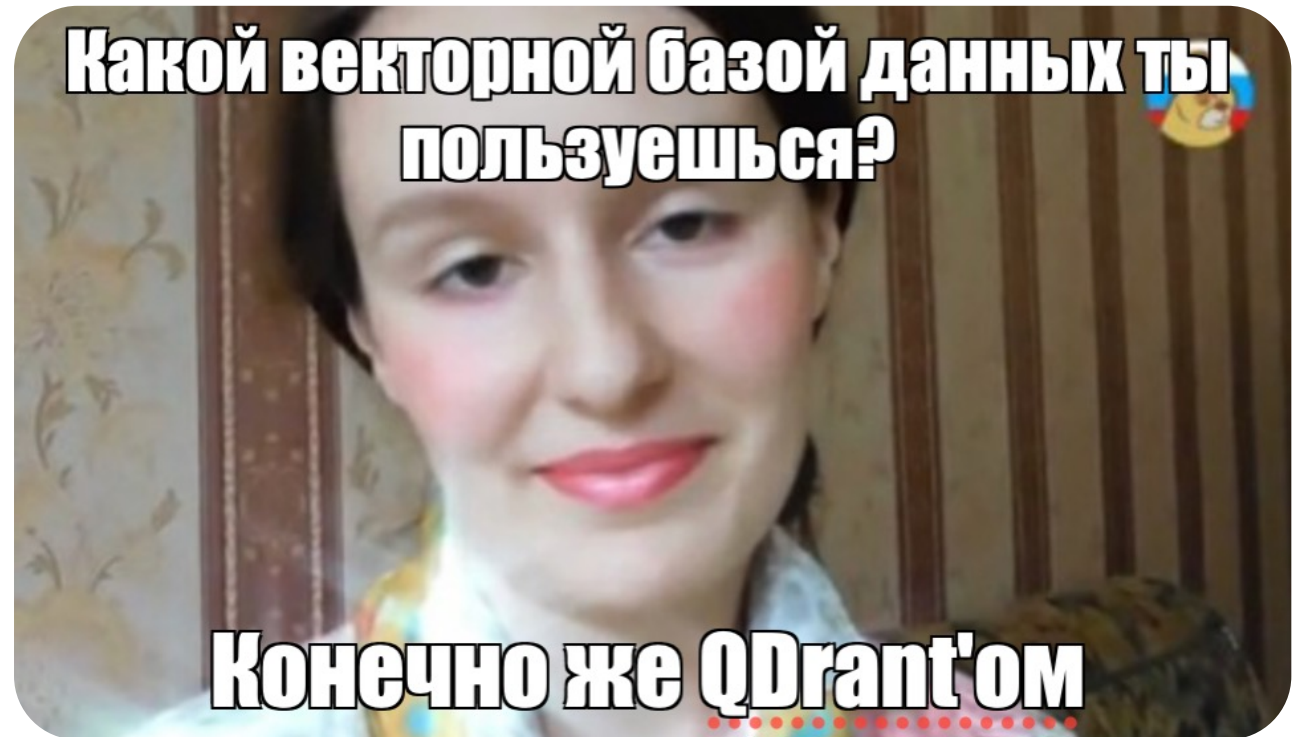


ЧТО ДЕЛАТЬ С РОСТОМ КОЛИЧЕСТВА ИСТОЧНИКОВ

точка банк

Как делать надо:

1. Использовать единую general-модель
2. Structure-aware chunking
3. Per-service index
4. Использовать метаданные!
5. Использовать инкрементальное обновление



ИТОГИ

точка банк

1. Улучшили качество
2. Улучшили производительность
3. Хорошо работаем с любым количеством документов



точка банк



**СПАСИБО
ЗА ВНИМАНИЕ!**

Влад Попов

 @YaZOp





.ML

Тг-канал нашей команды