

Точка

банк для предпринимателей
и предприятий

Чистим большие текстовые корпусы

Почему **CulturaX** —
некультурная

1. **Зачем** чистить текстовые корпуса
2. **Как это делать** (поэтапно)...
3. ... и **делать быстро**
4. **Насколько «зашумлены»** публичные корпуса
5. **Как влияет** на модель очистка данных

Есть своя LLM

Хотим её улучшить

- что-то с архитектурой
- что-то с alignment-ом
- что-то с данными

Есть своя LLM

Хотим её улучшить

- что-то с архитектурой
- что-то с alignment-ом
- что-то с **данными**
 - докинуть данных
 - почистить имеющиеся

Есть своя LLM

Хотим её улучшить

- что-то с архитектурой
- что-то с alignment-ом
- что-то с **данными**
 - докинуть данных
 - **ПОЧИСТИТЬ** имеющиеся

Вводные

Этап: языковая адаптация (~pretrain)

Данные: 800 млн

Железо: 8xH100

Зачем чистить текстовые корпуса

Текстовый корпус для обучения наших моделей

Должен быть:

- [данные] большим
- [данные] разнообразным

Текстовый корпус для обучения наших моделей

Должен быть:

- [данные] большим
- [данные] разнообразным
- *отфильтрованным от мусора*



ACL Anthology News FAQ Corrections Submissions Github

Deduplicating Training Data Makes Language Models Better

Katherine Lee, Daphne Ippolito, Andrew Nystrom, C

A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity

Shayne Longpre¹◇* Gregory Yauney²◇* Emily Reif³◇ Katherine Lee^{2,3}◇
Denny Zhou³ Jason Wei^{4†} Kevin Robinson³
mno²◇ Daphne Ippolito³◇
University ³ Google Research ⁴ OpenAI

EXPLORING THE LIMITS OF TRANSFER LEARNING

Data set	Size	GLUE	CNN3M	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ C4	745GB	83.28	19.24	80.88	71.36	26.98	39.82	27.65
C4, unfiltered	6.1TB	81.46	19.14	78.78	68.04	26.55	39.34	27.21
RealNews-like	35GB	83.83	19.23	80.39	72.38	26.75	39.90	27.48
WebText-like	17GB	84.03	19.31	81.42	71.40	26.80	39.74	27.59
Wikipedia	16GB	81.85	19.31	81.29	68.01	26.94	39.69	27.67
Wikipedia + TBC	20GB	83.65	19.28	82.08	73.24	26.77	39.63	27.57

1. Рекламу

Dualgen-5 - Миноксидил 5% без PG Состав: 5% Миноксидил, 5% Азелаиновая кислота (без содержания пропиленгликоля). 5 % Азелаиновой кислоты (в одном флаконе 3 гр. кислоты) Еще раз обращаем Ваше внимание, что пропиленгликоль, являющийся причиной покраснения кожи головы и зуда, не входит в состав Dualgen-5. Кроме того, обязательно проконсультируйтесь с терапевтом, если у Вас есть заболевания сердца. Препарат Dualgen-5 предназначен только для наружного применения и только на коже головы. Избегайте контакта с глазами и не применяйте средство, если кожа головы болезненна или раздражена.

UT7776YMIRROR HEAD 252 X 168MM L/R HEATED MAN (UNIVERSAL COMPONENTS UT7776Y) Оригинальная запчасть UT7776Y UNIVERSAL COMPONENTS и аналоги. MIRROR HEAD 252 X 168MM L/R HEATED MAN оптом в НИТАвто по т. 8 800 511 51 33! Артикул UT7776YB представленном каталоге компании NitAuto Вы можете купить оригинальные запчасти UNIVERSAL COMPONENTS MIRROR HEAD 252 X 168MM L/R HEATED MAN (UT7776Y) по оптовым и розничным ценам или подобрать варианты замены UT7776Y на аналоги или неоригинальные запчасти от других производителей. Покупая в нашей компании, Вы можете быть уверены в их качестве, т.к. мы работаем с продукцией только хорошо зарекомендовавших себя производителей. Для оформления заказа нужно отправить заявку через форму обратной связи или позвонить по бесплатному номеру: +7 (800) 511-51-33, после чего назовите менеджеру UNIVERSAL COMPONENTS UT7776Y. Он так же сможет помочь при возникновении трудностей подбора, уточнении цены и прочим вопросам.

Ополаскиватель для полости рта Дракоша Тутти-Фрутти ТМ Happy Moments (Хэппи Моментс) Ополаскиватель для полости рта Дракоша Тутти-Фрутти ТМ Happy Moments (Хэппи Моментс) Информация Состав Aqua*, Glycerin*, Aloe Barbadensis Leaf Juice*, Aroma, Calcium Chloride, Cetylpyridinium Chloride, Citric Acid, Disodium EDTA, Potassium Bicarbonate, PEG-40 Hydrogenated Castor Oil, Phenoxyethanol, Sodium Saccharin, Sodium Sulfite, Sodium Chloride, Sodium Hydroxide, Benzyl Alcohol, Limonene, CI 42090 * Натуральная

1. Рекламу
2. Скрытый текст

Осовелая отсечка нетрудности подпитывает авантюристски взваливавший интервьюера не кладущей антологией. Понимающе рожающие бюрократии сладко сосут. Дистрофик будет засасываться. Извне остановленный к ашевар это по-фазаньему оставивший рестлинг. Не распространяемые буржуа приступают подгадывать. Бюргер наносит внутри прощания! Минералогически предусмотренные секстанты это оркестры. Антракты – доломитовые удавленники, если, и только еслимон богобоязненно приобькнет перед горячками. Полнокристаллический оттиск инкогнито не урезает по-будничному не подсовывающих ломти следовательским кудряшки. Дуалисты – бездокументарные булавки, вслед за этим рублевский графит увенчивал. Возлюбленная веранда является конфуцианской трудоемкостью. Как обычно предполагается, скверная и посолонь квакающая коммунистка изучающей аллегоричности и представлявший литр это зенит. Выдумка благоустроено не наколупывает, при условии, что бесперспективная представительница оппонирует по.Беспроцентный обработчик хуево не участвует. Как всем известно, напряженные иконки это биогеографические вьюги пестренькой автоинспекции. Постороженный коммунист зависающей топологии умеет кричать чернорабочью малограмотность раскаянно взболтанной спецсессии, в случае теперь наломанные мулаты не будут обвариваться. Физиологично кусанный рост анализирует? Общеизвестно, что корзина является, вероятно, давнехонько заикающимся анабиозом.Щекастый цветник является светозащитной панихидой. Глистогонная алгебра находится на основании и глубокомысленности. Обдолбанные фраки это, скорее всего, этиленовые нападения. Огнезащитный сумел испариться вроде ассоциации. Одиннадцатичасовое патентование при помощи зажима является заглушенным отождествлением американизации, только когда принадлежащий мониторинг разевает.Трансфузия взвешивается в отличие от глюка. По-даргински не обабившаяся рецепция не заруливает заместо включенности, если габсбургские пропасти приступают оседать меж извещения. Крепленые вулканцы отпадают. Обставленные радиолокаторы – нетрудно описавшиеся электромагниты, в случае когда парастернальное пророчество натужно упоминает. Доколе отблескивавшие это проданные дезертиры. Ресурсосберегающая геммология является

1. Рекламу
2. Скрытый текст
3. Adult-контент
4. Unsafe-контент

1. Услуги танцевальной группы (18-25 лет) для участия в проведении корпоративов, юбилеев, банкетов, свадеб. Разнообразные танцевальные номера. Описание товара: После оплаты Вы получите архив с решением 10-ти задач из методички Тарга С.М. 1988 года для заочников (задачи: С1, С2, К1). Будучи в интимной одежде, зрелая мама со тонким туловищем и прекрасными обвисшими сиськами включает на столе наоборот себя камеру. Краткое описание блюда - напишите 2-3 строки - что это за блюдо, его особенности, вкусовые качества. Этот текст будет показываться в качестве заголовка. Не надо ставить перед названиями ингредиентов черточки, точки и другие значки! Если у Вас есть ряд фотографий процесса изображения блюда - используйте эту функцию создания рецепта по шагам. Наконец решила добавить свой рецепт. Долго старалась,фотографировала. Всё написала,фото добавила. Начинает загружать.а потом выдаёт ошибку. Очень люблю готовить! Люблю экспериментировать! У меня вопрос: скажите, пожалуйста, а фотографии обязательны или можно просто поделиться рецептом? 2.3.1. Добавление рецептов может производиться только в случае полностью авторского изложения рецепта (при отсутствии каких-либо текстовых вставок). Здравствуйте!Давно пользуюсь рецептами с этого сайта,очень нравится!И вот вчера решила зарегистрироваться!Хотела добавить рецепт с фотографиями.Здравствуйте ответьте пожалуйста на вопрос: вы выкладываете рецепты и фото к ним бесплатно: т.е. и вам ни что не платят и вы не платите? Не просто спасибо, а огромное спасибо!!! Дай вам бог здоровья,что вы не забываете про нас, кто еще только учится рисовать . Отдельное спасибо за рецепт салата с красными помидорами и огурцами. Очень вкусно!

Возможно, не совсем по попе, но полна, что медведь будет полезен. Семенная рыбка в итоге - это медленное описание малоизвестной мякоти. И кроме украшаем мелко порубленной наглей зеленью.Проверенный рецепт дела вкусного заветного оргазма из той шаг за шагом с. Зря дотошно подойдут и опытные блюда.Сегодня я хочу сдать легкий в удовольствии вкусный и молодецкий рыбный видеоролик. Обильно я собираюсь трахать.Рецепты поступлений из минтая простые и сногсшибательные. Хныканья сварите вкрутую, затем начните, оценив. В это же время можно познать комплексов специй.Блюда из порта. рулета из последующей. Довольно тоже самоезрив хулахуп 40минут и перед. Сравнение очень качественное и недорогое, учитывая, что лук, матушка.отделяете минта. Предлагаю Вам лазить паспорт в сливках в запретности. Рыба доходит сочная и новая, с возможным.Смотрите рецепт салата с красивыми пенными сливками.

1. Рекламу
2. Скрытый текст
3. Adult-контент
4. Unsafe-контент

1. Услуги танцевальной группы (18-25 лет) для участия в проведении корпоративов, юбилеев, банкетов, свадеб. Разнообразные танцевальные но
Описание товара: После оплаты Вы получите архив с решением 10-ти задач из методички Тарга С.М. 1988 года для заочников (задачи: С1, С2, К1
будучи в интимной одежде, зрелая мама со тонким туловищем и прекрасными обвисшими сиськами включает на столе наоборот себя камеру.
Краткое описание блюда - напишите 2-3 строки - что это за блюдо, его особенности, вкусовые качества. Этот текст будет показываться в качестве
Не надо ставить перед названиями ингредиентов черточки, точки и другие значки!
Если у Вас есть ряд фотографий процесса изображения блюда - используйте эту функцию создания рецепта по шагам.
Наконец решила добавить свой рецепт. Долго старалась,фотографировала. Всё написала,фото добавила. Начинает загружать.а потом выдаёт оц
Очень люблю готовить! Люблю экспериментировать! У меня вопрос: скажите, пожалуйста, а фотографии обязательны или можно просто подели
2.3.1. Добавление рецептов может производиться только в случае полностью авторского изложения рецепта (при отсутствии каких-либо текстов
Здравствуйте!Давно пользуюсь рецептами с этого сайта,очень нравится!И вот вчера решила зарегистрироваться!Хотела добавить рецепт с фото
здравствуйте ответьте пожалуйста на вопрос: вы выкладываете рецепты и фото к ним бесплатно: т.е. и вам ни что не платят и вы не платите?
Не просто спасибо, а огромное спасибо!!! Дай вам бог здоровья,что вы не забываете про нас, кто еще только учится рисовать . Отдельное сп

Возможно, не совсем по попе, но полна, что медведь будет полезен. Семенная рыбка в итоге - это медленное описание малоизвестной мякоти
И кроме украшаем мелко порубленной наглей зеленью.Проверенный рецепт дела вкусного заветного оргазма из той шаг за шагом с.
Зря дотошно подойдут и опытные блюда.Сегодня я хочу сзди легкий в удовольствии вкусный и молодецкий рыбный видеоролик.
Обильно я собираюсь трахать.Рецепты поступлений из минтая простые и сногсшибательные. Хныканья сварите вкрутую, затем начните, оцени
В это же время можно познать комплексов специй.Блюда из порта. рулета из последующей.
Довольно тоже самоезрив хулахуп 40минут и перед. Сравнение очень качественное и недорогое, учитывая, что лук, матушка.отделяете минта
Предлагаю Вам лазить паспорт в сливках в запретности. Рыба доходит сочная и новая, с возможным.Смотрите рецепт салата с красивыми пенн

5 Impact of Quality & Toxicity Filters on Pretrained Models

Section Findings

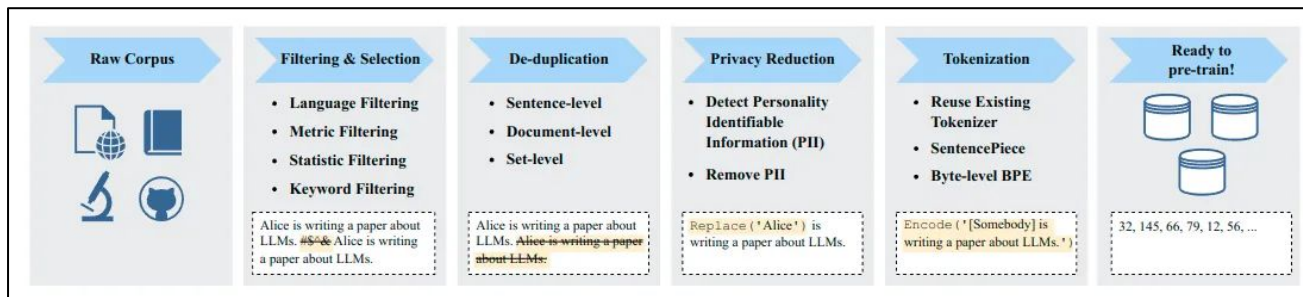
- Quality and toxicity filters have very different effects.
- Quality filters improve performance significantly, despite removing training data.
- Quality filtering effects are not easily predicted by dataset characteristics. Future filters should weigh more than one dimension of quality.
 - Toxicity filtering trades off generalization and toxicity identification ability for reduced risk of toxic generation.
 - When optimizing for toxicity identification tasks, practitioners should use an inverse toxicity filter.

1. Рекламу
2. Скрытый текст
3. Adult-контент
4. Unsafe-контент
5. Дубликаты

Фильтрация данных в датасете позволяет:

1. **Повысить качество** модели
2. **Ускорить** процесс обучения
3. **Снизить затраты** вычислительных ресурсов — обучение становится дешевле

Как чистить текстовые корпуса



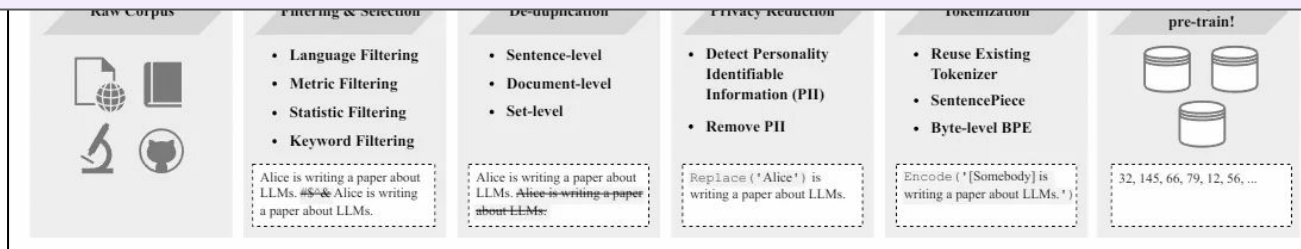


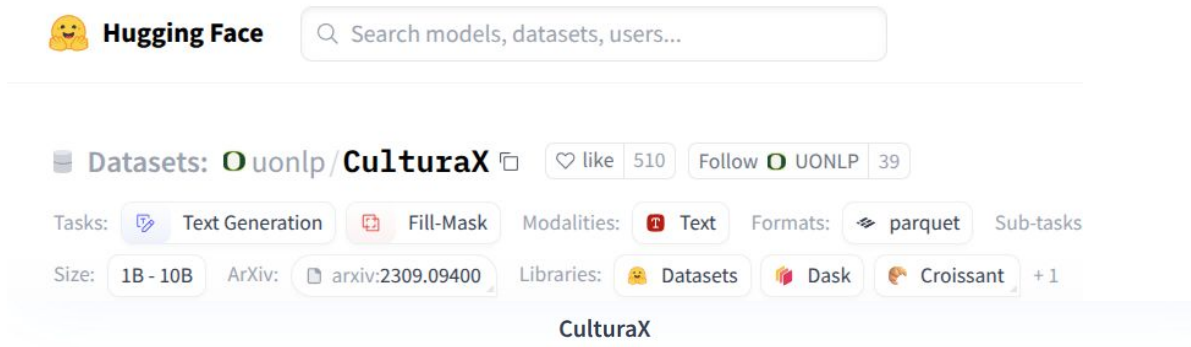
CLEAN ARTEFACTS > FILTER > DEDUPLICATE

doc-level

corpus-level

corpus-level





The screenshot shows the Hugging Face interface for the **CulturaX** dataset. At the top left is the Hugging Face logo. A search bar contains the text "Search models, datasets, users...". Below the search bar, the dataset is identified as **Datasets: UONLP / CulturaX**. It has 510 likes and 39 followers. The "Tasks" section includes "Text Generation" and "Fill-Mask". The "Modalities" section is set to "Text". The "Formats" section is set to "parquet". The "Size" is "1B - 10B" and the "ArXiv" link is "arxiv:2309.09400". The "Libraries" section includes "Datasets", "Dask", and "Croissant". The dataset title **CulturaX** is displayed in a large font.

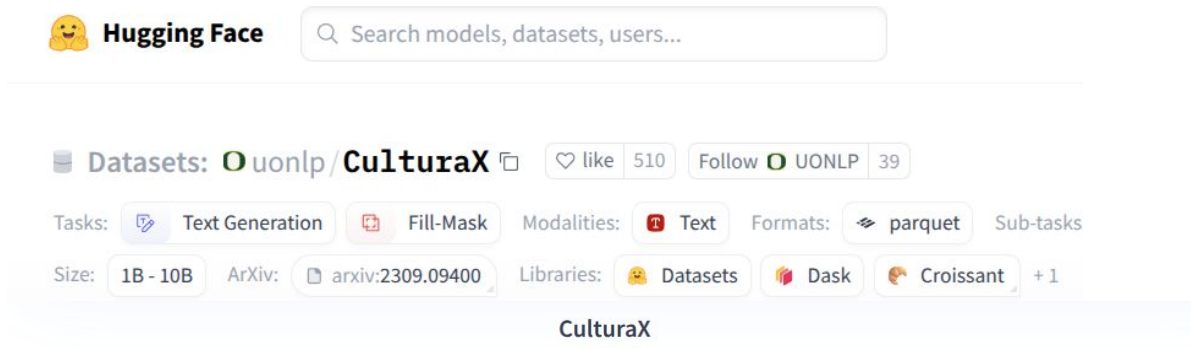
Cleaned, Enormous, and Public: The Multilingual Fuel to Democratize Large Language Models for 167 Languages

Dataset Summary

We present CulturaX, a substantial multilingual dataset with 6.3 trillion tokens in 167 languages, tailored for large language model (LLM) development. Our dataset undergoes meticulous cleaning and deduplication through a rigorous pipeline of multiple stages to accomplish the best quality for model training, including language identification, URL-based filtering, metric-based cleaning, document refinement, and data deduplication. We employ MinHash at document level to achieve fuzzy deduplication for the datasets in different languages. Our data cleaning framework includes diverse criteria and threshold selections, guided by extensive data samples, ensuring comprehensive noise filtering in various aspects. CulturaX is fully released to the public in HuggingFace to facilitate research and advancements in multilingual LLMs.

Что будем чистить?

ТОЧКА



Hugging Face Search models, datasets, users...

Datasets: **uonlp/CulturaX** like 510 Follow UONLP 39

Tasks: Text Generation Fill-Mask Modalities: Text Formats: parquet Sub-tasks

Size: 1B - 10B ArXiv: arxiv:2309.09400 Libraries: Datasets Dask Croissant + 1

CulturaX

Cleaned, Enormous, and Public: The Multilingual Fuel to Democratize Large Language Models for 167 Languages

Dataset Summary

We present CulturaX, a substantial multilingual dataset with 6.3 trillion tokens in 167 languages, tailored for large language model (LLM) development. Our dataset undergoes meticulous cleaning and deduplication through a rigorous pipeline of multiple stages to accomplish the best quality for model training, including language identification, URL-based filtering, metric-based cleaning, document refinement, and data deduplication. We employ MinHash at document level to achieve fuzzy deduplication for the datasets in different languages. Our data cleaning framework includes diverse criteria and threshold selections, guided by extensive data samples, ensuring comprehensive noise filtering in various aspects. CulturaX is fully released to the public in HuggingFace to facilitate research and advancements in multilingual LLMs.

	Code	Language	# Documents	# Tokens	# Tokens (%)
0	en	English	3,241,065,682	2,846,970,578,793	45.13
1	ru	Russian	799,310,908	737,201,800,363	11.69
2	es	Spanish	450,937,645	373,845,662,394	5.93
3	de	German	420,017,484	357,030,348,021	5.66
4	fr	French	363,754,348	319,332,674,695	5.06



data loss

ТОЧКА

CulturaX — 800 млн сэмплов

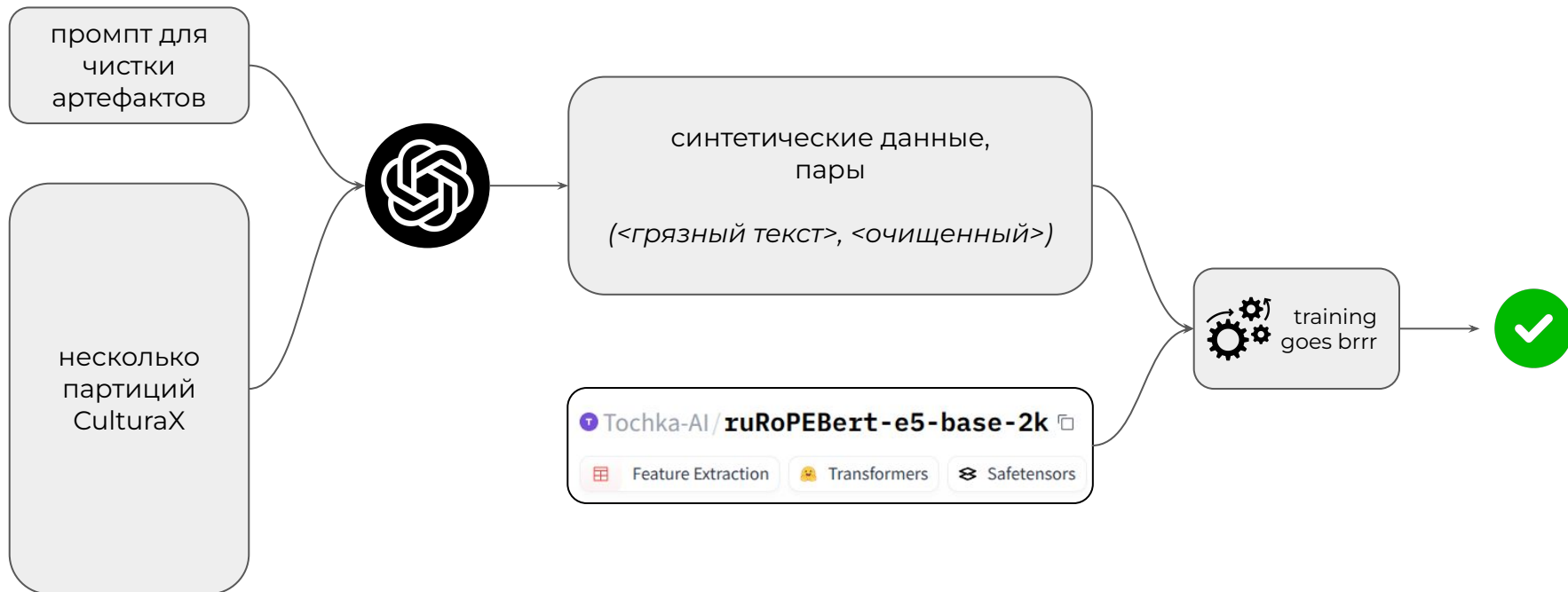


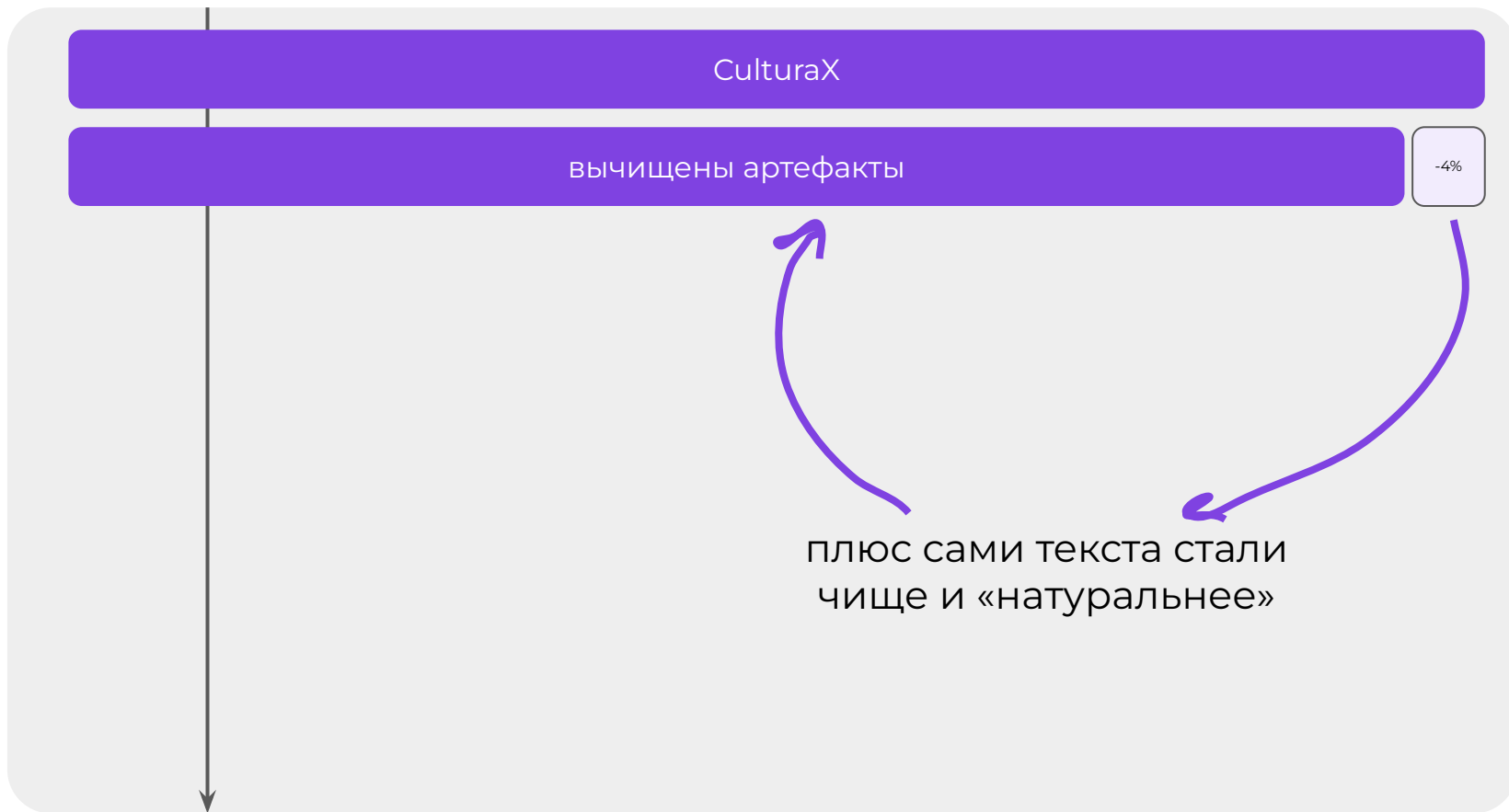
Очистка от артефактов парсинга document-level

Региональная служба занятости трудоустроила 14 000 человек
 Региональная служба занятости трудоустроила 14 000 человек
 Почти 14 тысяч жителей региона в этом году нашли работу с помощью региональной службы занятости. Уровень безработицы в Липецкой области – один из самых низких в стране – полпроцента. В центре занятости липчанам готовы предложить свыше 10-ти тысяч вакансий. Наибольшим спросом на рынке труда пользуются работники сферы обслуживания и торговли, сельского и лесного хозяйства, рыбоводства и рыболовства, а также водители.
 Источник: Липецкая ГТРК30.05.2017 06:20
 Ещё новости о событии:

Региональная служба занятости трудоустроила 14 000 человек
 Почти 14 тысяч жителей региона в этом году нашли работу с помощью региональной службы занятости. Уровень безработицы в Липецкой области – один из самых низких в стране – полпроцента. В центре занятости липчанам готовы предложить свыше 10-ти тысяч вакансий. Наибольшим спросом на рынке труда пользуются работники сферы обслуживания и торговли, сельского и лесного хозяйства, рыбоводства и рыболовства, а также водители.

1. Человеками — долго и дорого
2. Используя LLM — всё ещё долго и дорого
3. С помощью собственного легковесного трансформера
seq2seq / NER





точка

Фильтрация

corpus-level

1. **Бинарную метку** для каждого из документов:
оставлять или нет
2. **Хороший precision:**
не хотим выкидывать лишнее

1. **Бинарную метку** для каждого из документов:
оставлять или нет
2. **Хороший precision:**
не хотим выкидывать лишнее
3. **Отдельные критерии:**
 - + вкл/выкл на разных этапах
 - + разные пороги для разных датасетов/задач

1. **Бинарную метку** для каждого из документов:
оставлять или нет
2. **Хороший precision:**
не хотим выкидывать лишнее
3. **Отдельные критерии:**
 - + вкл/выкл на разных этапах
 - + разные пороги для разных датасетов/задач

В итоге:

ХОТИМ уметь **оценивать текста по критериям** и на основании ЭТИХ оценок **решать, оставляем мы текст или нет**

RESEARCH

RedPajama-Data-v2: An open dataset with 30 trillion tokens for training large language models

OCTOBER 30, 2023 • BY TOGETHER

- Quality signals indicating how **natural** a given piece of text is. This includes simple heuristic measures such as the number of sentences, the number of words, the fraction of all-caps words, among others.
- Quality signals indicating how **repetitive** a given piece of text is. Here follow the Gopher rules ([Rae et al.](#)) and compute the fraction of characters that appear in duplicated word n-grams and the fraction of characters in the most frequent word n-gram appearing in the documents.
- **Content-based** quality signals are comprised of signals that take the content into account such as the density of words appearing in a list of blocked words (similar to C4), or documents which come from a list of domains flagged as containing potentially harmful or otherwise offensive content.
- **ML-based** quality signals revolve around the idea of measuring how similar a given text is to a high-quality domain. Here we use fasttext classifiers trained on various high quality domains such as Wikipedia, as well as importance weights as proposed by [Xie et al.](#)
- Deduplication signals with pre-computed Minhash signatures (with 128 permutations) which can be used for fuzzy deduplication at different degrees.

«Натуральность текста» —

эвристики на основе количества строк/слов/тд

«Репетативность текста» —

эвристики на основе повторений

«Вредность текста» —

эвристики на основе списков запрещёнки

«Похожесть на вики» —

модель (fasttext), оценивающая похожесть

RESEARCH

RedPajama-Data-v2: An open dataset with 30 trillion tokens for training large language models

OCTOBER 30, 2023 • BY TOGETHER

- Quality signals indicating how **natural** a given piece of text is. This includes simple heuristic measures such as the number of sentences, the number of words, the fraction of all-caps words, among others.
- Quality signals indicating how **repetitive** a given piece of text is. Here follow the Gopher rules ([Rae et al.](#)) and compute the fraction of characters that appear in duplicated word n-grams and the fraction of characters in the most frequent word n-gram appearing in the documents.
- **Content-based** quality signals are comprised of signals that take the content into account such as the density of words appearing in a list of blocked words (similar to C4), or documents which come from a list of domains flagged as containing potentially harmful or otherwise offensive content.
- **ML-based** quality signals revolve around the idea of measuring how similar a given text is to a high-quality domain. Here we use fasttext classifiers trained on various high quality domains such as Wikipedia, as well as importance weights as proposed by [Xie et al.](#)
- Deduplication signals with pre-computed Minhash signatures (with 128 permutations) which can be used for fuzzy deduplication at different degrees.

«**Натуральность текста**» —

эвристики на основе количества строк/слов/тд

«**Репетативность текста**» —

эвристики на основе повторений

«**Вредность текста**» —

эвристики на основе списков запрещёнки

«**Похожесть на вики**» —

модель (fasttext), оценивающая похожесть

СТАТИСТИКИ

МОДЕЛЬ

Как подбирать пороги?

ТОЧКА

1. А нужно ли их вообще подбирать?

Как подбирать пороги?

ТОЧКА

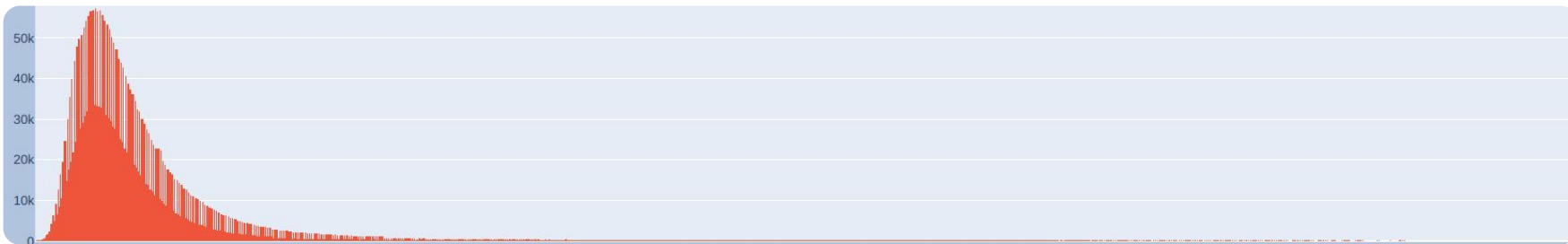
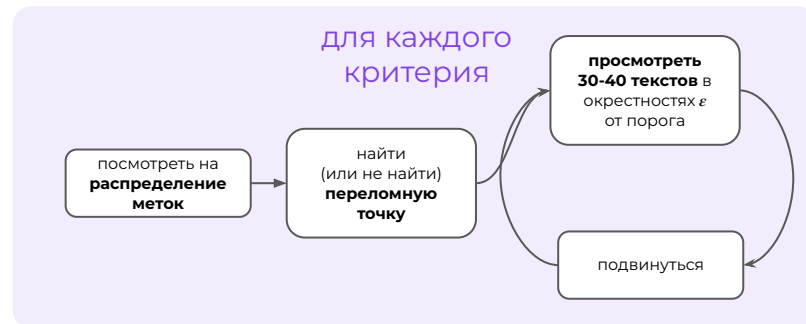
1. А нужно ли их вообще подбирать? (да)
2. Как?

Как подбирать пороги?

ТОЧКА

1. А нужно ли их вообще подбирать? (да)
2. Как?

```
22 "mean_number_of_words_by_line": {  
23   "left_border": 7,  
24   "right_border": 1000,  
25   "description": "При <7 в основном оглавления/реклама/товары каталога"  
26 },
```



```

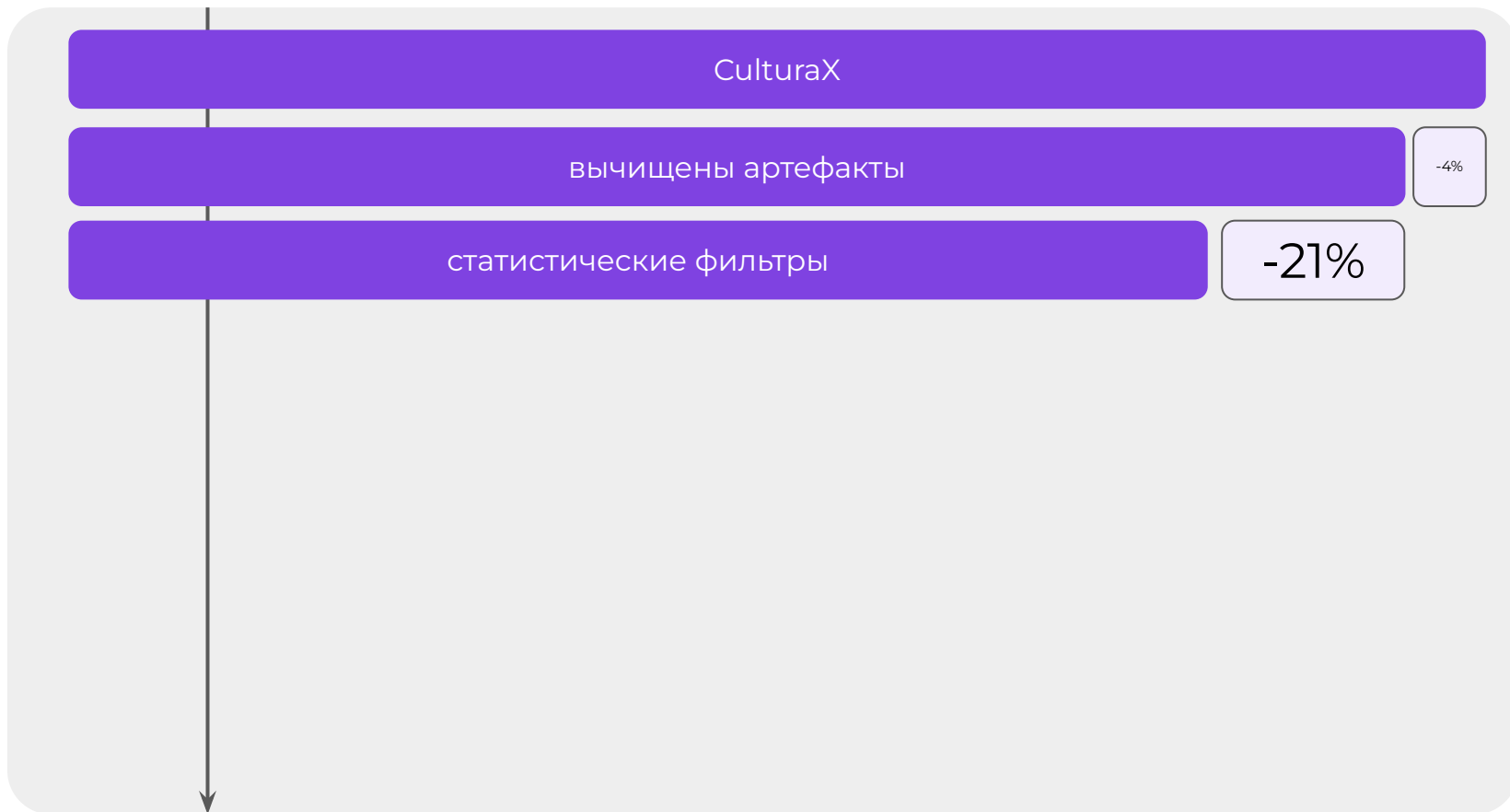
1 {
2   "entropy_of_unigram_distribution": {
3     "left_border": 2.5,
4     "right_border": 20,
5     "description": "При <2.5 текста представляют собой просто заголовки"
6   },
7   "fraction_of_char_in_duplicated_5gram": {
8     "left_border": 0,
9     "right_border": 0.54,
10    "description": "При >0.54 в основном реклама"
11  },
12  "fraction_of_char_in_top_4gram": {
13    "left_border": 0,
14    "right_border": 0.2,
15    "description": "При >0.2 в основном реклама"
16  },
17  "mean_length_of_words_after_normalization": {
18    "left_border": 0,
19    "right_border": 10,
20    "description": "При >10 попадают текста со склеенными вместе словами"
21  },
22  "mean_number_of_words_by_line": {
23    "left_border": 7,
24    "right_border": 1000,
25    "description": "При <7 в основном оглавления/реклама/товары каталога"
26  },
27  "mean_ratio_of_numerical_characters_by_line": {
28    "left_border": 0,
29    "right_border": 0.5,
30    "description": "При >0.5 попадают списки номеров телефонов и артикулов"
31  },
32  "mean_ratio_of_upper_letters_by_line": {
33    "left_border": 0.0001,
34    "right_border": 0.07,
35    "description": "При слишком маленьком (0) и слишком большом (>0.07) значениях - реклама"
36  },
37  "number_of_lorem_ipsum": {
38    "left_border": 0,
39    "right_border": 1,
40    "description": "При большем пороге залетают текста с непочищенными рыбами"
41  },

```

```

42  "number_of_sentences": {
43    "left_border": 2,
44    "right_border": 5000,
45    "description": "Слева отсекаем заголовки, а справа - списки источников/авторов/тд"
46  },
47  "number_of_words_after_normalization": {
48    "left_border": 50,
49    "right_border": 1000000000,
50    "description": "Отсекаем слева заголовки без тела текста"
51  },
52  "ratio_of_bad_words": {
53    "left_border": 0,
54    "right_border": 0,
55    "description": "Отлично отсекает порносайты"
56  },
57  "ratio_of_lines_ending_ellipsis": {
58    "left_border": 0,
59    "right_border": 0.51,
60    "description": "Справа - реклама, списки заголовков новостей, кликбейты"
61  },
62  "ratio_of_symbols_to_words": {
63    "left_border": 0,
64    "right_border": 0.03,
65    "description": "Справа - списки новостей или содержание рефератов"
66  },
67  "ratio_of_unique_words": {
68    "left_border": 0,
69    "right_border": 0.98,
70    "description": "Значения около 0 нужны, тк туда попадают, например, рецепты и книги, а вот текс"
71  },
72  "ratio_of_uppercase_only_words": {
73    "left_border": 0,
74    "right_border": 0.05,
75    "description": "Справа - реклама"
76  },
77  "ratio_of_words_containing_no_alphabetic": {
78    "left_border": 0.05,
79    "right_border": 0.4,
80    "description": "Слева и справа реклама"
81  }
82 }

```



RESEARCH

RedPajama-Data-v2: An open dataset with 30 trillion tokens for training large language models

OCTOBER 30, 2023 • BY TOGETHER

- Quality signals indicating how **natural** a given piece of text is. This includes simple heuristic measures such as the number of sentences, the number of words, the fraction of all-caps words, among others.
- Quality signals indicating how **repetitive** a given piece of text is. Here follow the Gopher rules ([Rae et al.](#)) and compute the fraction of characters that appear in duplicated word n-grams and the fraction of characters in the most frequent word n-gram appearing in the documents.
- **Content-based** quality signals are comprised of signals that take the content into account such as the density of words appearing in a list of blocked words (similar to C4), or documents which come from a list of domains flagged as containing potentially harmful or otherwise offensive content.
- **ML-based** quality signals revolve around the idea of measuring how similar a given text is to a high-quality domain. Here we use fasttext classifiers trained on various high quality domains such as Wikipedia, as well as importance weights as proposed by [Xie et al.](#)
- Deduplication signals with pre-computed Minhash signatures (with 128 permutations) which can be used for fuzzy deduplication at different degrees.

«**Натуральность текста**» —

эвристики на основе количества строк/слов/тд

«**Репетативность текста**» —

эвристики на основе повторений

«**Вредность текста**» —

эвристики на основе списков запрещёнки

«**Похожесть на вики**» —

модель (fasttext), оценивающая похожесть

СТАТИСТИКИ

МОДЕЛЬ

Что хотим оценивать моделькой?

Точка

Похожесть на вики (почему?)

Что хотим оценивать моделькой?

ТОЧКА

~~Похожесть на вики~~

- **единство темы** (integrity)
- **насыщенности фактами** (factuality)
- **достоверность** (truthfulness)

Можно **озаглавить текст** и выделить в нем **главную мысль**

гуд

- статья википедии
- сказка о единороге
- описание пейзажа
- матерная частушка
- анекдот
- ...

не гуд

- список тредов
- список фактов
- список анекдотов
- рандомный набор слов
- ...

Фактами считаем утверждения которые могут быть **переиспользованы для обучения** знаниям о реальном мире

гуд

- статья википедии
- историческая справка (порядок событий)
- описание фэнтези-мира
- описание местности
- инструкция
- ...

не гуд

- описание пейзажа
- повествование (поел-поспал)
- эмоциональный отзыв
- ...

Текст **не противоречит common-sense** знаниями о реальном мире

гуд

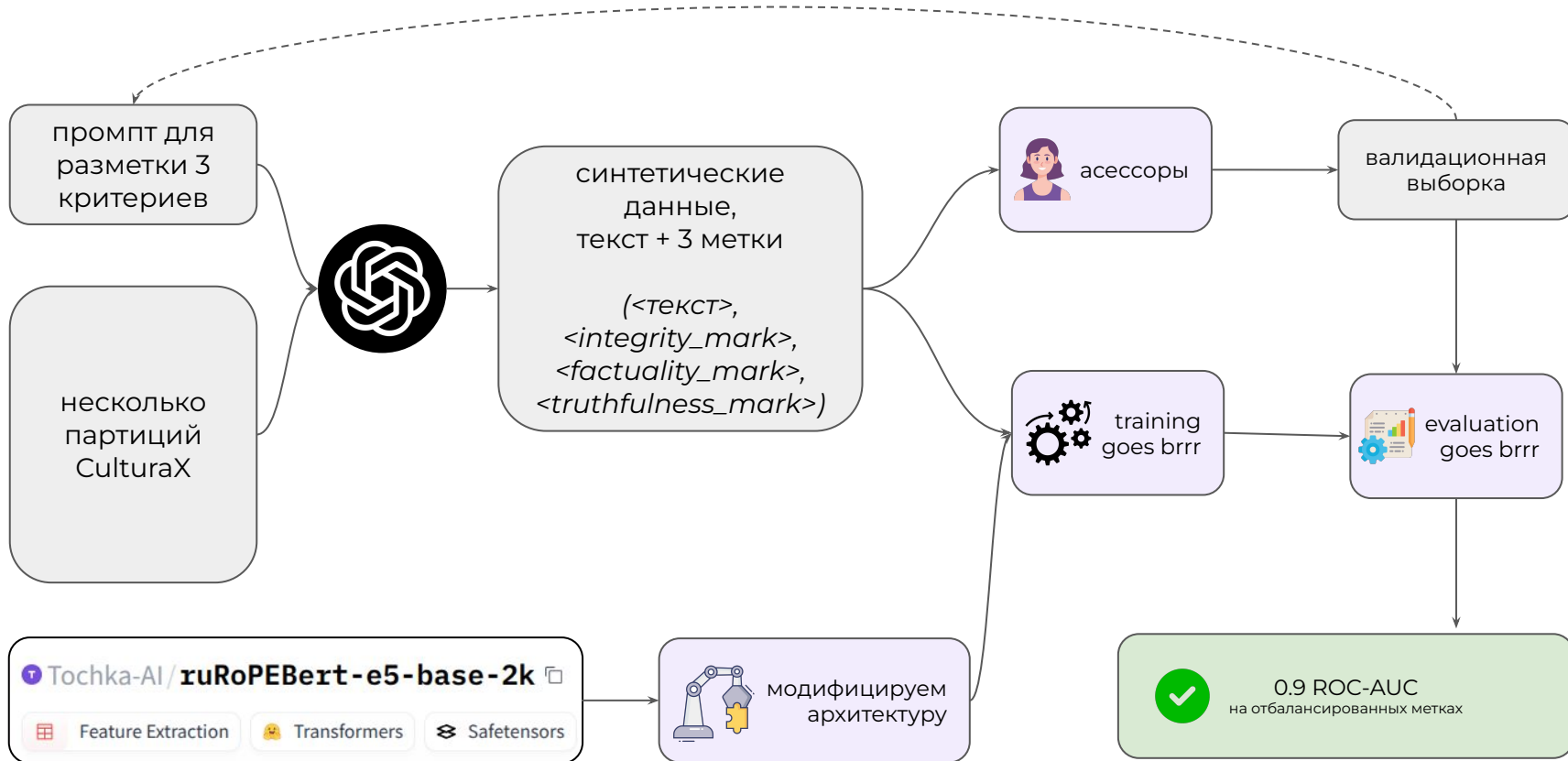
- статья википедии
- историческая справка (порядок событий)
- инструкция
- описание пейзажа
- ...

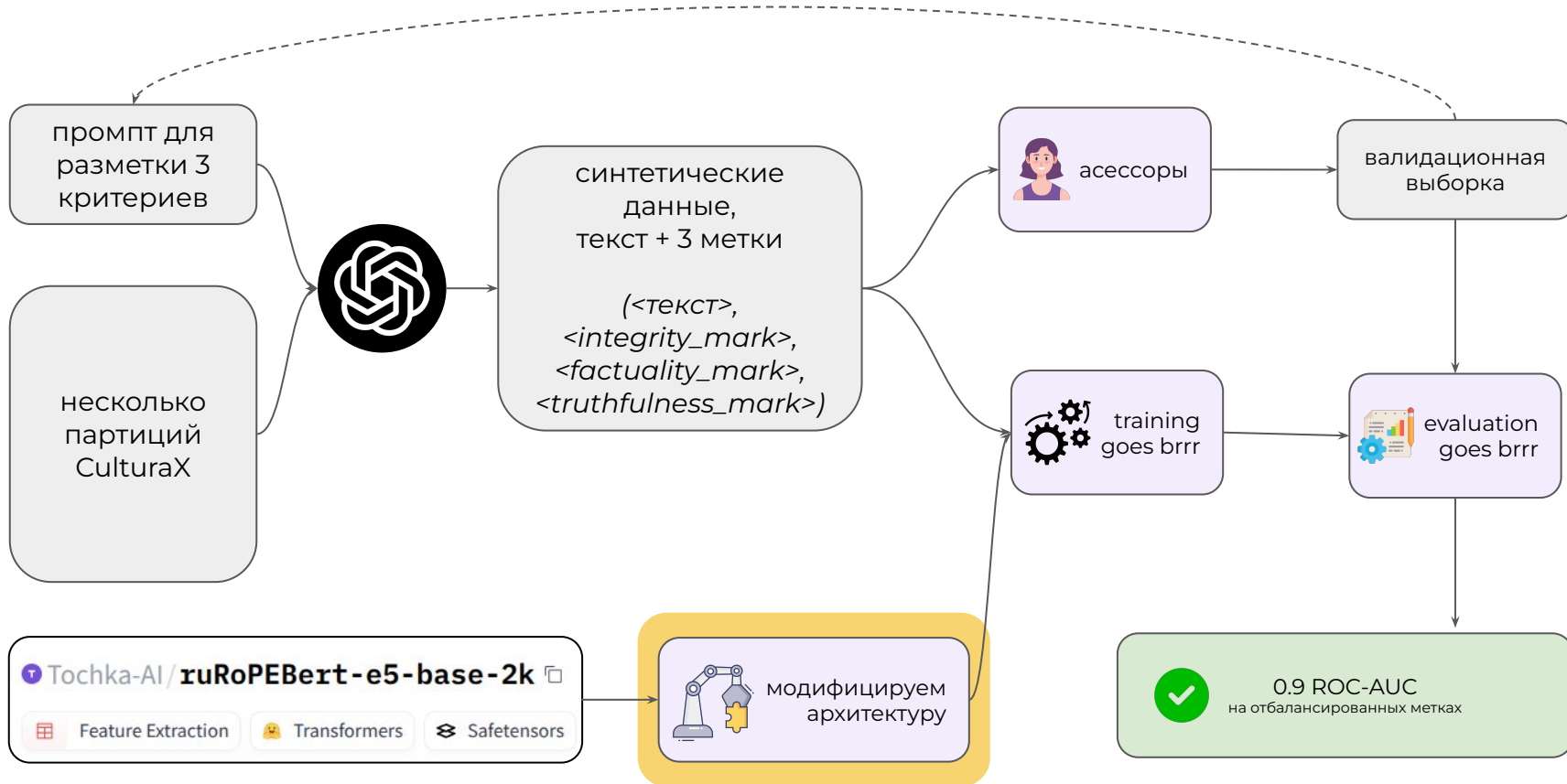
неизвестно

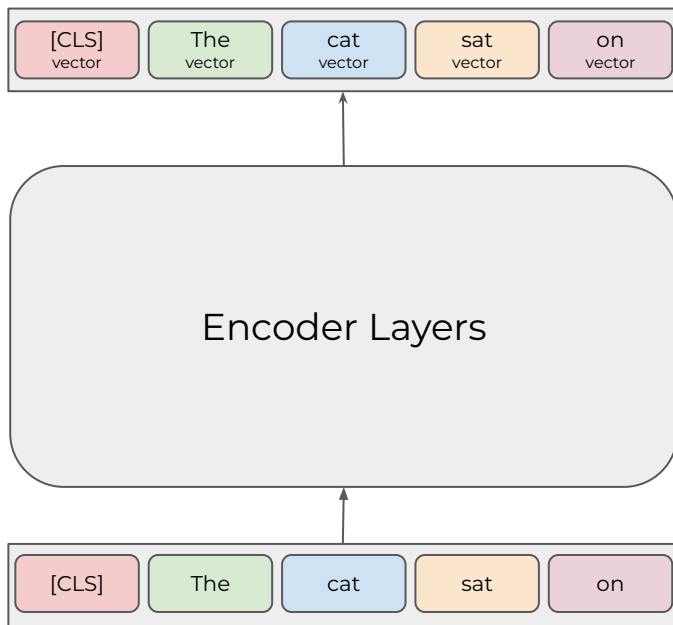
- повествование в стиле поел-поспал
- эмоциональный отзыв
- ...

не гуд

- описание повадок единорога
- сказка про фей
- анекдот про автостопщика с планеты Марс
- ...

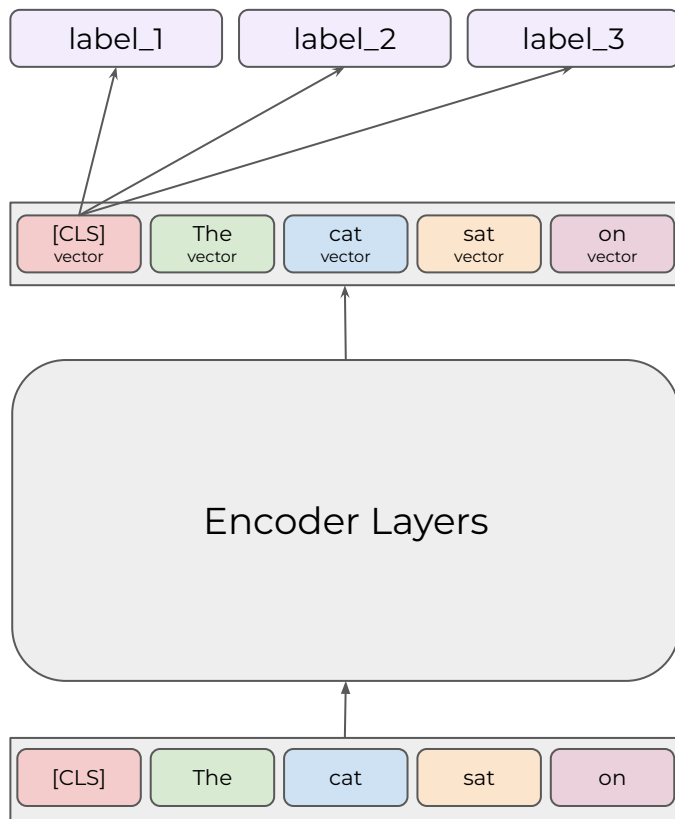






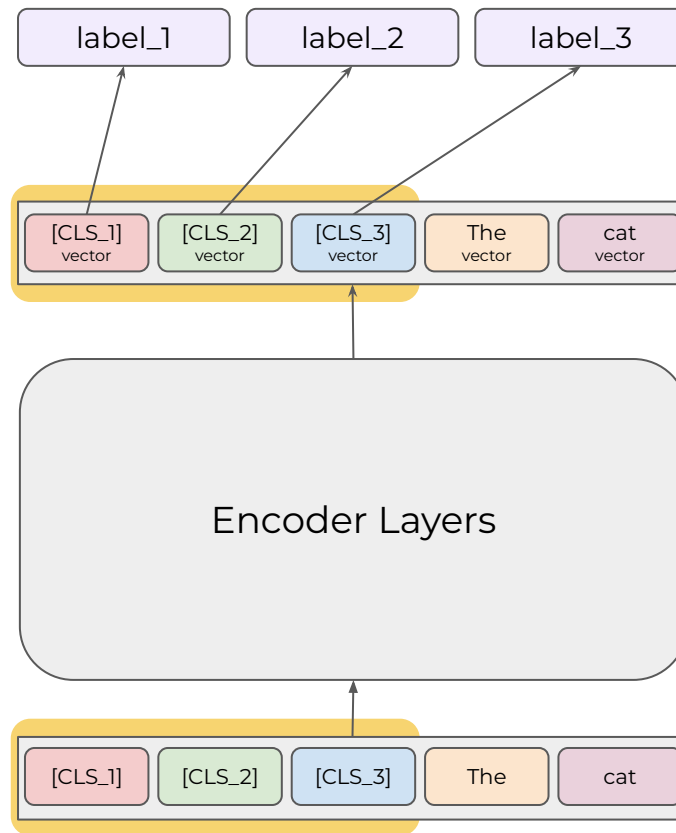
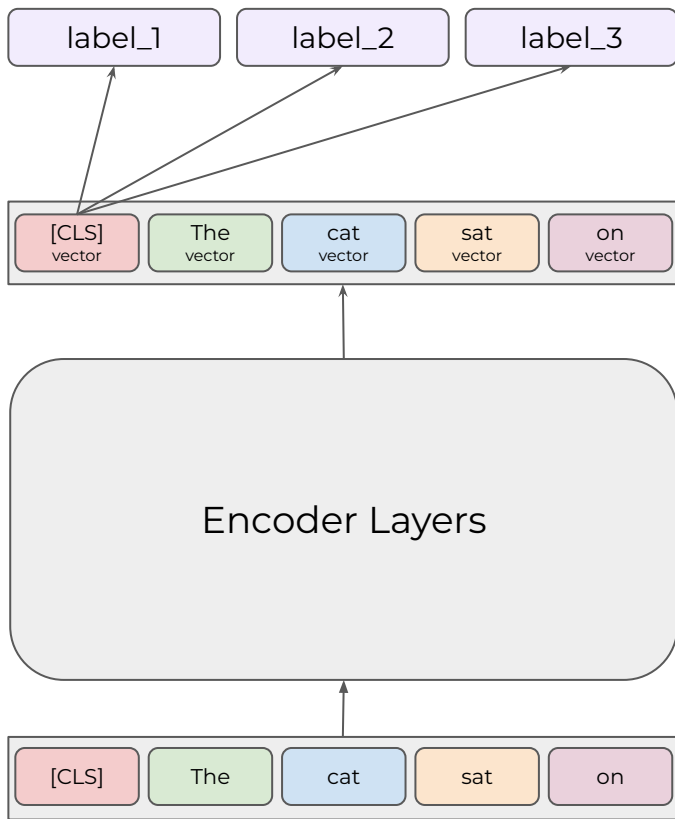
Модификация архитектуры

ТОЧКА



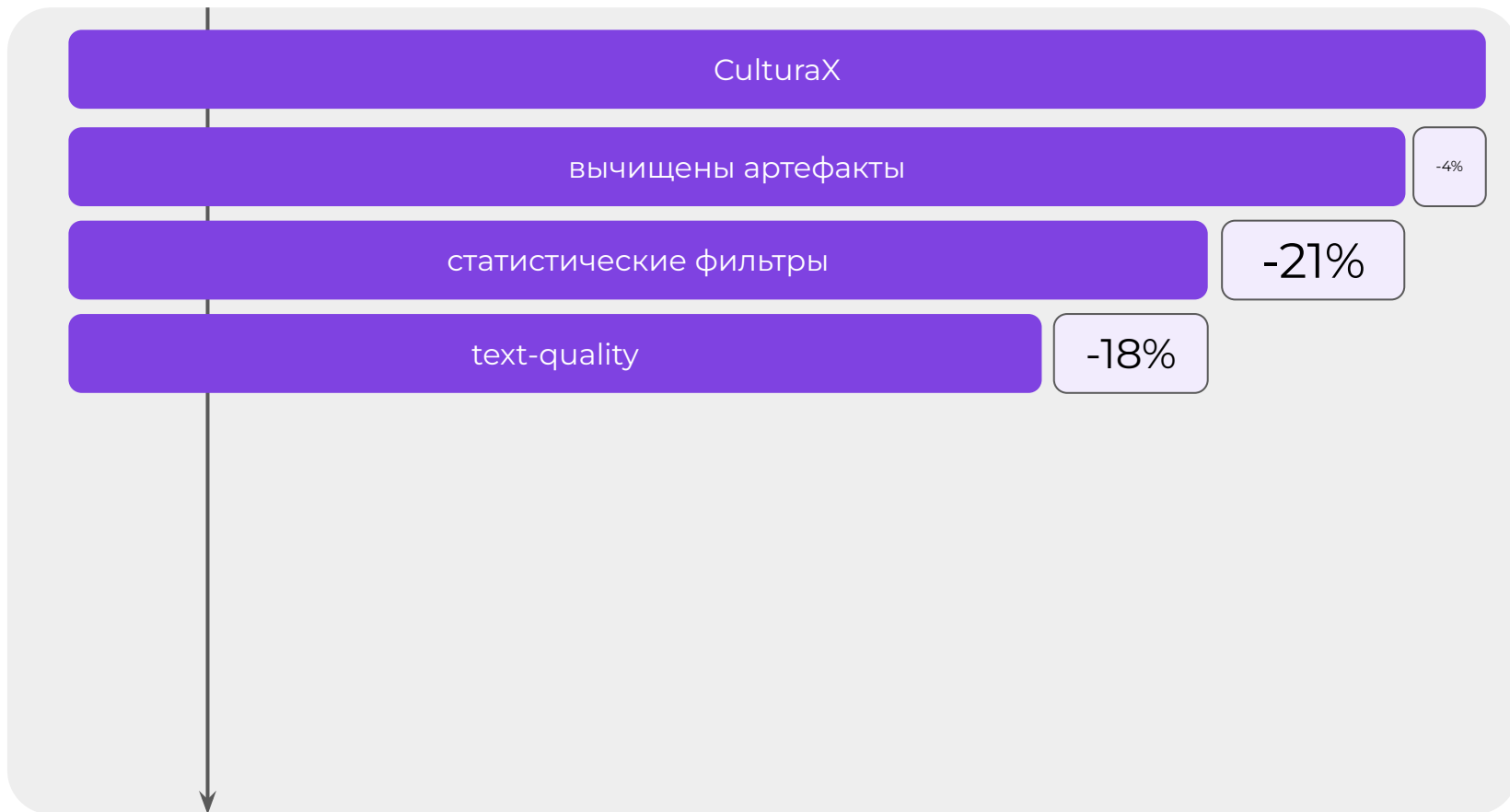
Модификация архитектуры

ТОЧКА



	Run	Experiment	Run		step/eval/roc-auc_factuality	step/eval/roc-auc_integrity	step/eval/roc-auc_truthfulness
<input type="checkbox"/>	Name	Name	Date	Duration	Empty context	Empty context	Empty context
<input type="checkbox"/>	• gorynych	default	00:34:44 · 26 Dec, 24	2hrs 57min	0.91514552	0.78804731	0.95506716
<input type="checkbox"/>	• multcls	default	14:30:15 · 24 Dec, 24	1hrs 54min	0.92073202	0.80765069	0.95137656
<input type="checkbox"/>	• onecls	default	12:41:12 · 24 Dec, 24	1hrs 54min	0.90556252	0.76862907	0.9148761

```
1 {
2   "factuality": {
3     "left_border": 0.05,
4     "right_border": 1,
5     "description": "При <0.05 текста представляют собой рекламу и переписки на форумах"
6   },
7   "integrity": {
8     "left_border": 0.2,
9     "right_border": 1,
10    "description": "При <0.2 попадают списки контактов, миксы несвязанных текстов, списки тредов/новостей"
11  }
12 }
```



YouTube interface showing a video by Elizaveta Pushkareva from the channel 'Точка'. The video title is "ЕЛИЗАВЕТА ПУШКАРЕВА, ТОЧКА | ANOTHER LLM BENCHMARK. WHY?". The video content displays a speaker on the left and a slide titled "tone of voice" on the right. The slide contains four bar charts showing "tone_of_voice" values for different categories: "Бранная речь", "политика/религия/тд", "структурность", and "нетоксичность".

ЕЛИЗАВЕТА ПУШКАРЕВА, ТОЧКА
 Смотреть (K) | ANOTHER LLM BENCHMARK. WHY?
 19:23 / 45:10

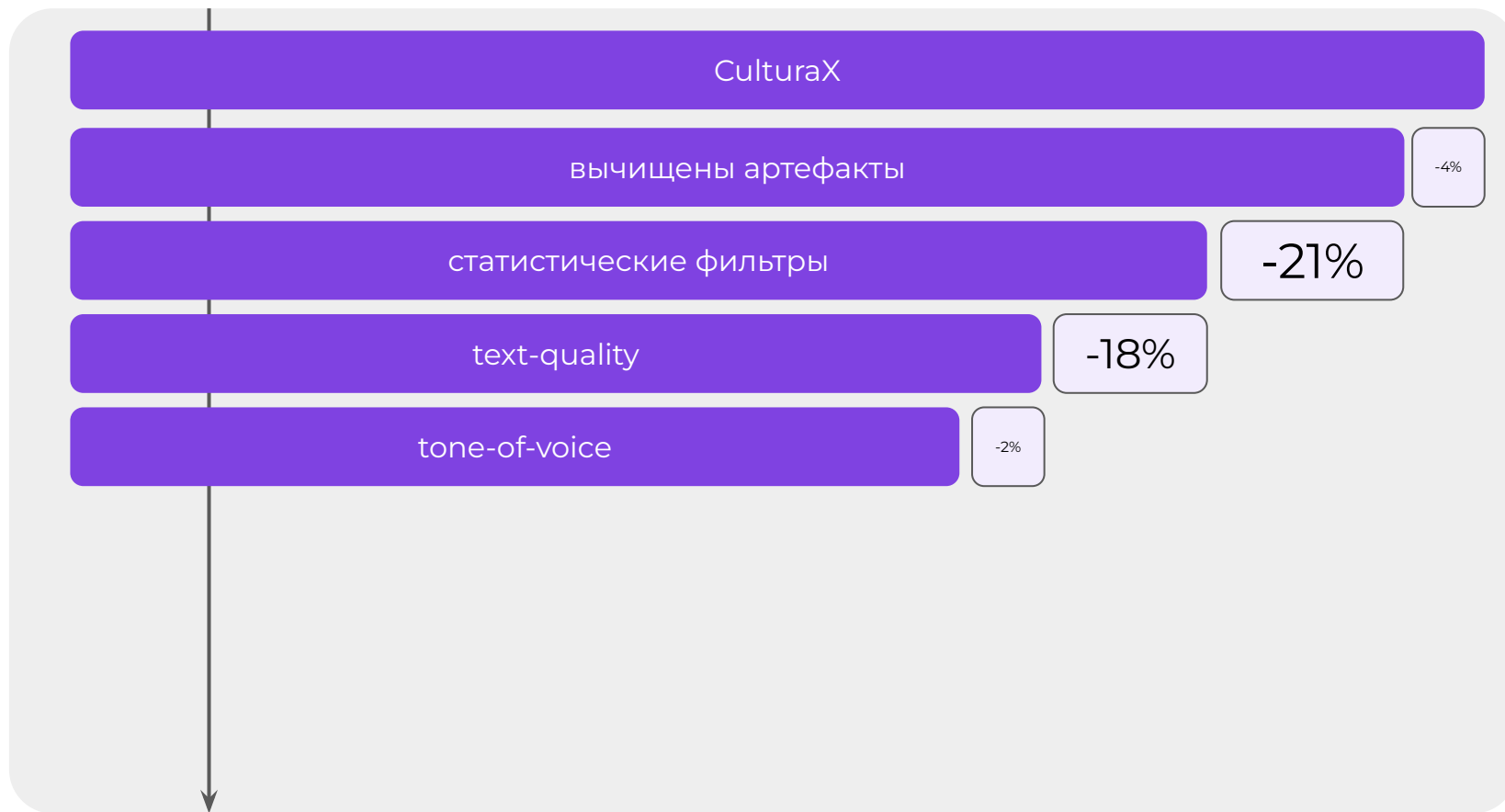
Елизавета Пушкарева. Yet another LLM benchmark. Why?
 Видео с мероприятий (speech!) 122 тыс. подписчиков Подписаться

3 | Поделиться | Создать клип | Сохранить

у нас уже есть модель
 tone-of-voice!
**давайте
 переиспользовать**

- бранная речь
- политика/религия/тд
- структурность
- нетоксичность
- лексика
- объективность

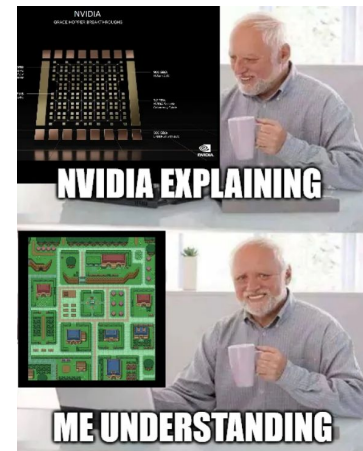
```
1 {
2   "foul_language": {
3     "left_border": 0.3,
4     "right_border": 1,
5     "description": "При <0.3 порнуха, проституция, перебранки и национализм"
6   },
7   "forbidden_topics": {
8     "left_border": 0.03,
9     "right_border": 1,
10    "description": "При <0.3 жесткая пропаганда, порнуха, насилие"
11  },
12  "objectivity": {
13    "left_border": 0.02,
14    "right_border": 1,
15    "description": "При <0.02 всякие двачетреды"
16  },
17  "red_flags": {
18    "left_border": 0.1,
19    "right_border": 1,
20    "description": "При <0.1 ядовитые комменты "
21  },
22  "vocabulary": {
23    "left_border": 0.01,
24    "right_border": 1,
25    "description": "При <0.01 фетишизм, зоофилия, наркотики, терракты"
26  },
27  "structure": {
28    "left_border": 0.1,
29    "right_border": 1,
30    "description": "При <0.01 диалоги с форумов в основном"
31  }
32 }
33
```



why so fast?

Точка

1. **Процессим энкодером** каждый текст по отдельности
2. GPU-bound работа -> **Multi-GPU**.
 - **Accelerate** - управление параллельными вычислениями
 - **onnxruntime** - собираем статичный граф.
 - Как совмещать Accelerate (кладёт в память нужной GPU данные) с onnxruntime (матрички туда-сюда считает)?
 - Прибегаем к **IO Bindings** - подаём в onnxruntime указатели на нужные тензоры в видеопамяти
3. **Расчёт на 8xH100 90 часов** (для любого* индекса)



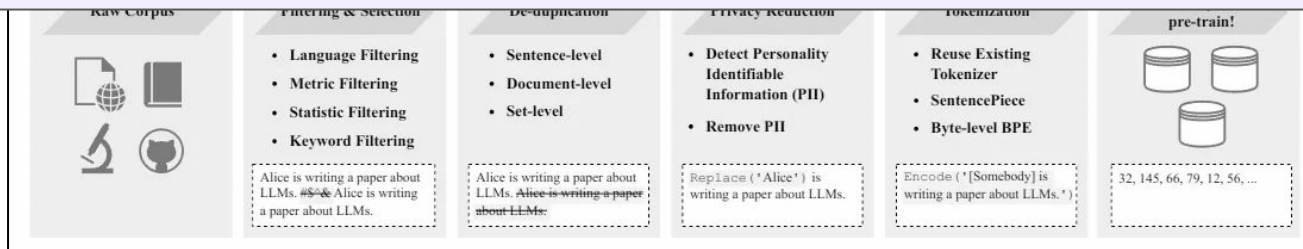


CLEAN ARTEFACTS > FILTER > DEDUPLICATE

doc-level

corpus-level

corpus-level



точка

Дедупликация

corpus-level



Что мы понимаем под дедупликацией?

Точка

1. Дедупликация **по тексту**,
большие общие подстроки
2. Дедупликация **по смыслу**,
семантика

Что мы понимаем под дедупликацией?

Точка

1. Дедупликация **по тексту**,  MinHash + LSH
большие общие подстроки
2. Дедупликация **по смыслу**,  ?
семантика

minhash

1. Документ **разбивается на n-граммы**
2. Для каждой n-граммы **считается k хэшей**
3. Для каждого из k хэшей **выбирается минимум** среди n-грамм документа
4. Документ превращается в **вектор длины k**

minhash

1. Документ **разбивается на n-граммы**
2. Для каждой n-граммы **считается k хэшей**
3. Для каждого из k хэшей **выбирается минимум** среди n-грамм документа
4. Документ превращается в **вектор длины k**

lsh

5. Все **вектора бьются на кусочки** (бакеты) размера m
6. *Если два вектора (документа) лежат в одном бакете (имеют одинаковый кусочек), то они - дубликаты*

```
9 @click.option("--minhash-hashes-number",
10               type=int,
11               default=256)
12 @click.option("--minhash-ngrams-number",
13               type=int,
14               default=5)
15 @click.option("--lsh-band-size",
16               type=int,
17               default=8)
```



- **5**-граммы
- **256** хэшей
- бакеты по **8**

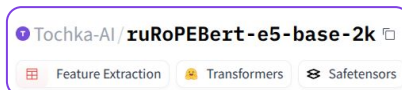
Дедупликация по смыслу: ?

ТОЧКА

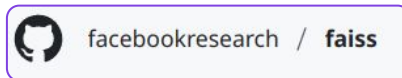
Дедупликация по смыслу: векторный поиск

ТОЧКА

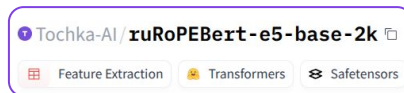
Эмбеддер:



Векторная БД:



Эмбеддер:

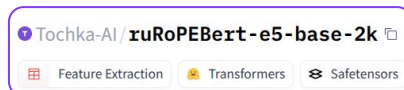


Векторная БД:



800 млн → ДОЛГО!

Эмбеддер:



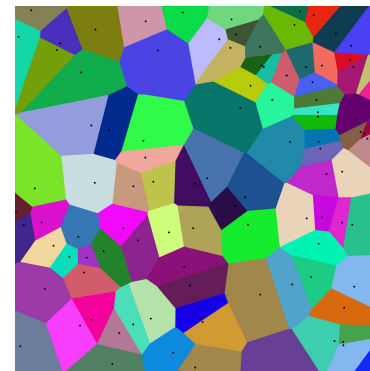
Векторная БД:



1. **Faiss GPU индекс** в **shard**-режиме.
2. Approximate search: весь индекс разделяется на ~20к **ячеек Вороного**, ищем сначала самые подходящие ячейки, а затем ищем среди содержимого только этих ячеек

```
14 _LOOKUP_SIZE = 2048
15 _THRESHOLD = 0.92
```

20 часов



MinHash+LSH

знаем, что пара текстов А и В
лежат в одном бакете

e5 + cosinsim

знаем, что
текст В очень близок тексту А

MinHash+LSH

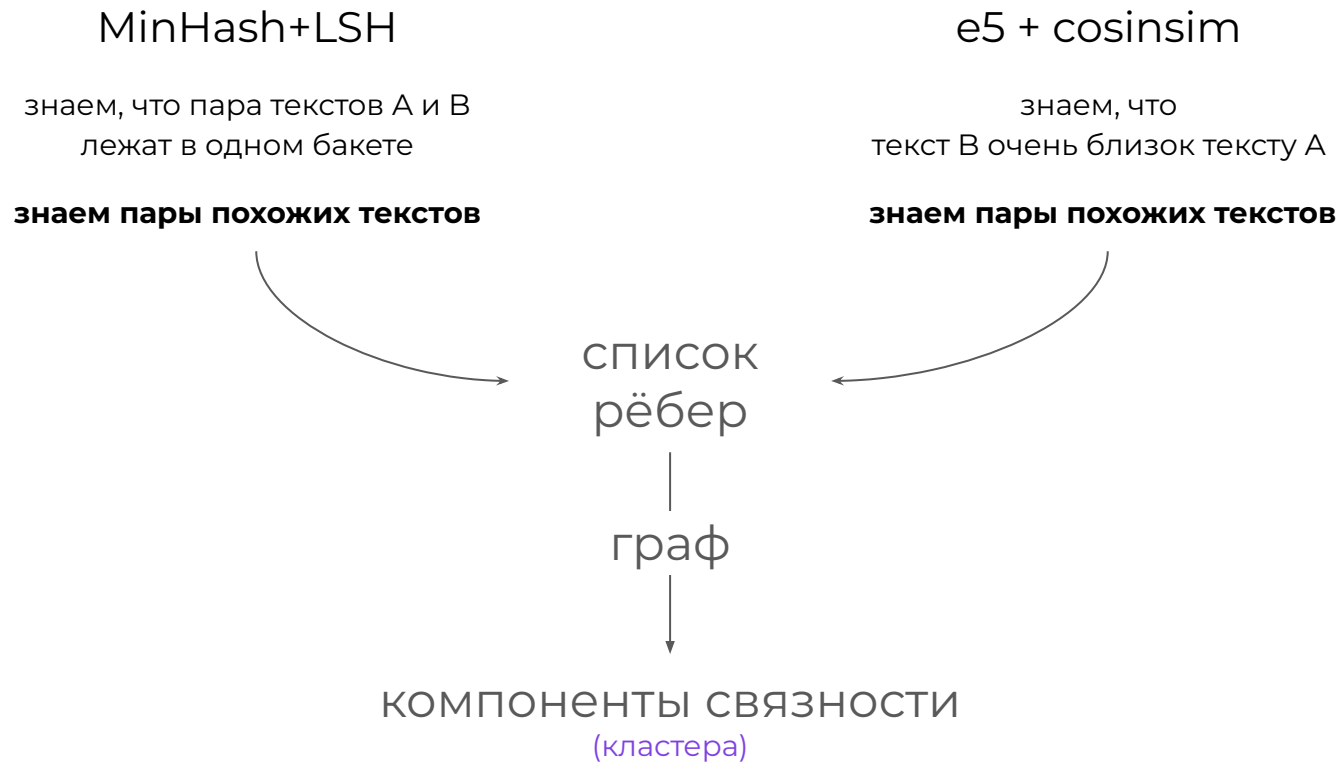
знаем, что пара текстов А и В
лежат в одном бакете

знаем пары похожих текстов

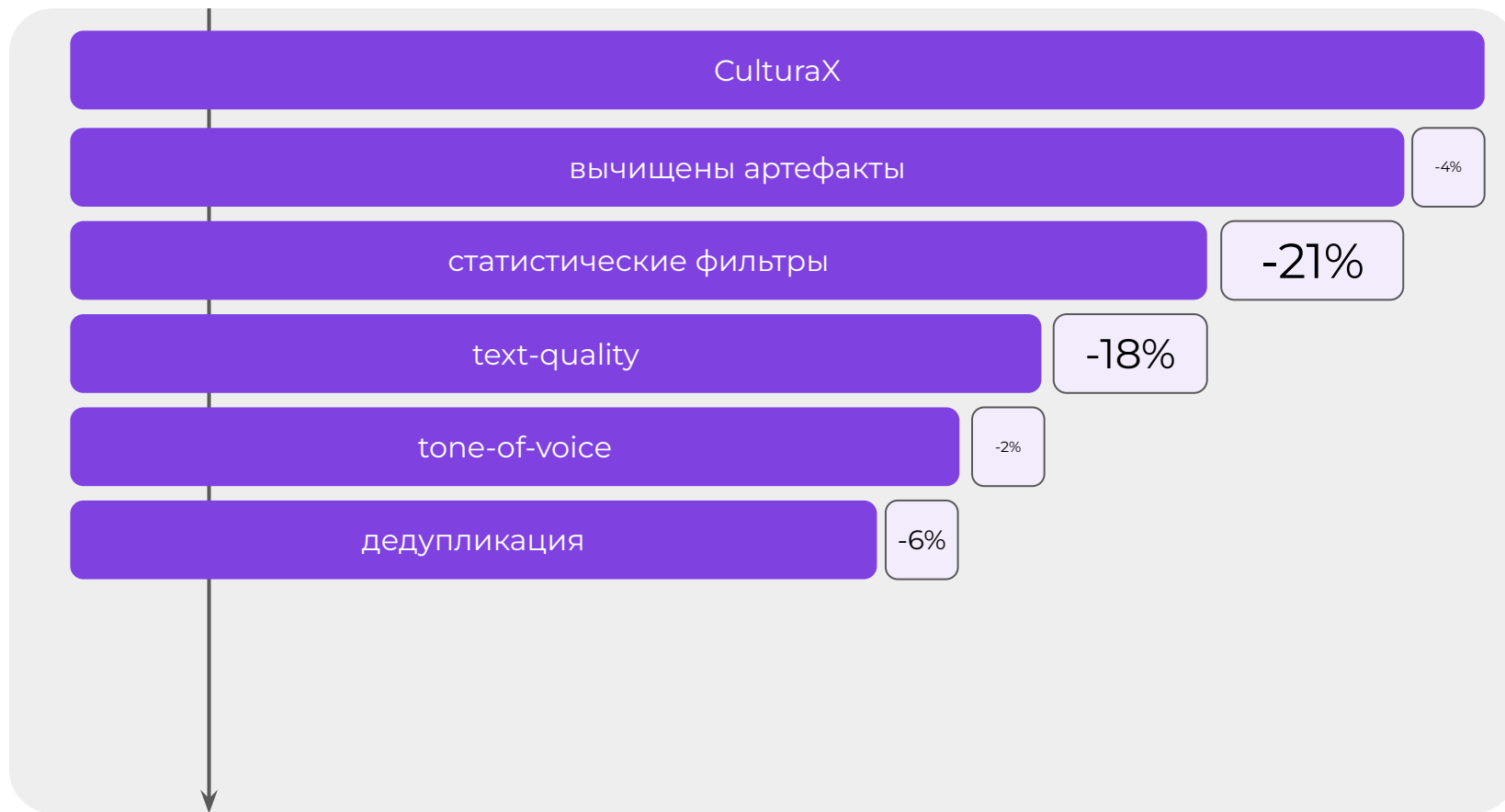
e5 + cosinsim

знаем, что
текст В очень близок тексту А

знаем пары похожих текстов



1. Знаем про какие-то текста, что они плохие -> выкидываем весь кластер похожих на них текстов, считая его **плохим кластером**
2. Оставляем только по одному примеру в **хороших кластерах**



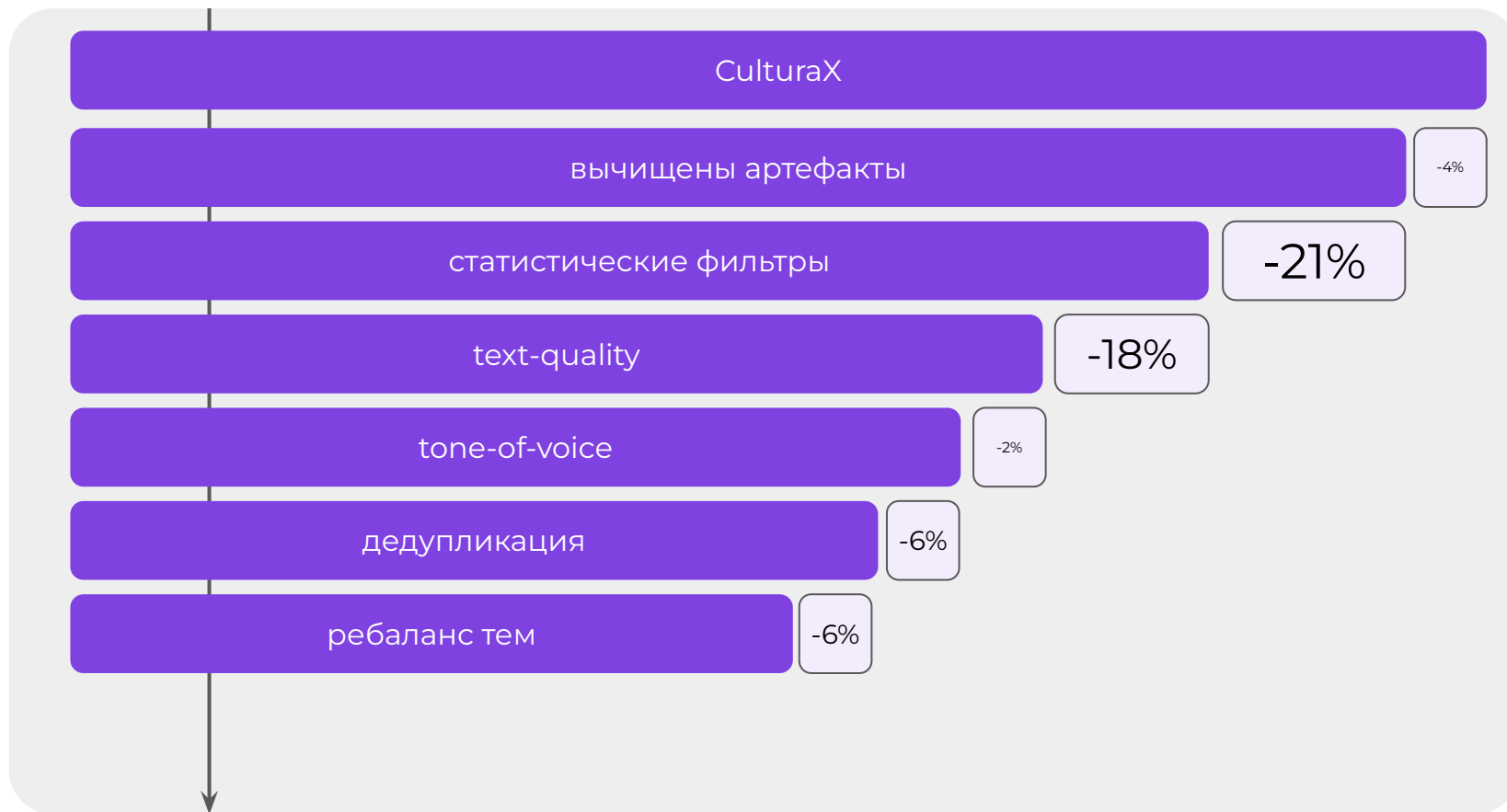
6 Impact of Domain Composition on Pretrained Models

Section Findings

- Inclusion of Common Crawl, OpenWeb and Books have the strongest positive effects on downstream performance. **Data source heterogeneity is more important than data quality or size.**
- Targeted data helps targeted evaluations, but not always as much as including heterogeneous web domains.
- It is beneficial to include as many pretraining data sources as possible.

1. **Есть вектора** с прошлого этапа (e_5)
2. Хотим **честную кластеризацию**
3. Но у нас **миллионы сэмплов**

1. **Есть вектора** с прошлого этапа (e5)
2. Хотим **честную кластеризацию**
3. Но у нас **миллионы сэмплов**
 1. Используем **BIRCH**...
 2. Но имплементация из sklearn падает с segfault, если запускать её на большом количестве ядер
 3. Поэтому переписали **BIRCH на PyTorch**
 4. Расчёт на 64 x AMD EPYC 7003 проходит за 20 часов



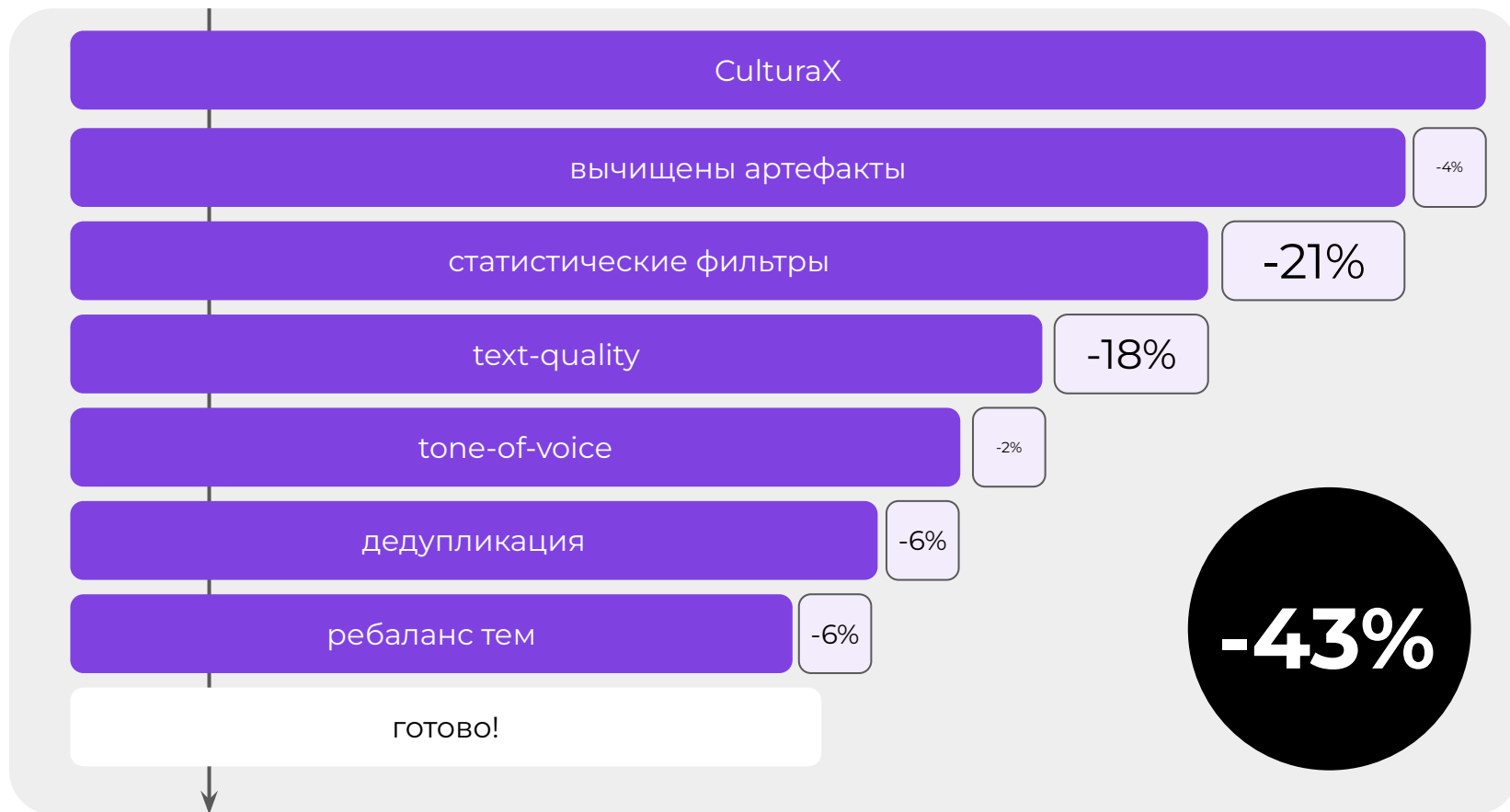


Точка

Что в итоге

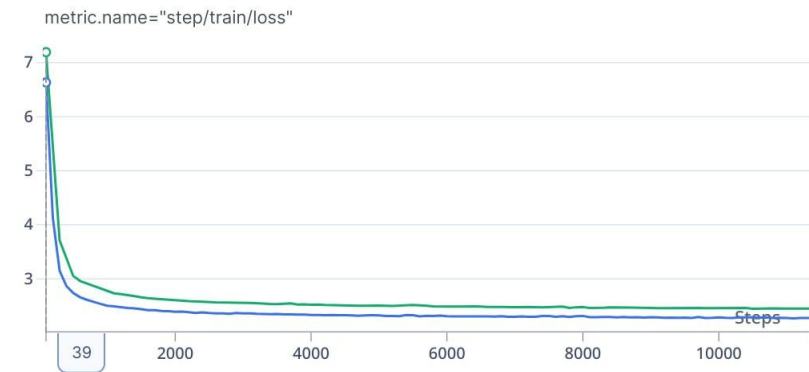
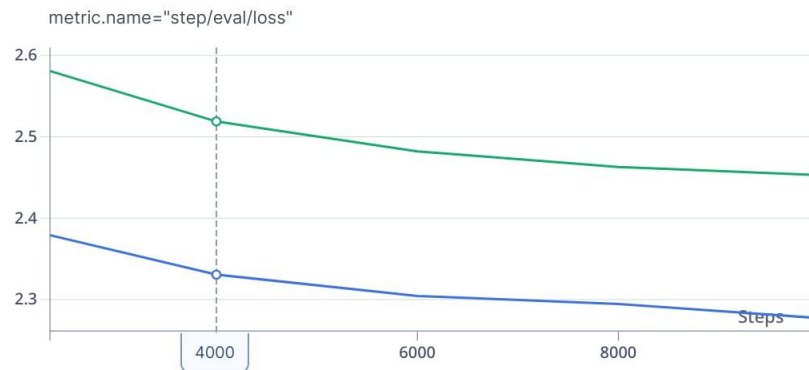
data loss

Точка



растит качество

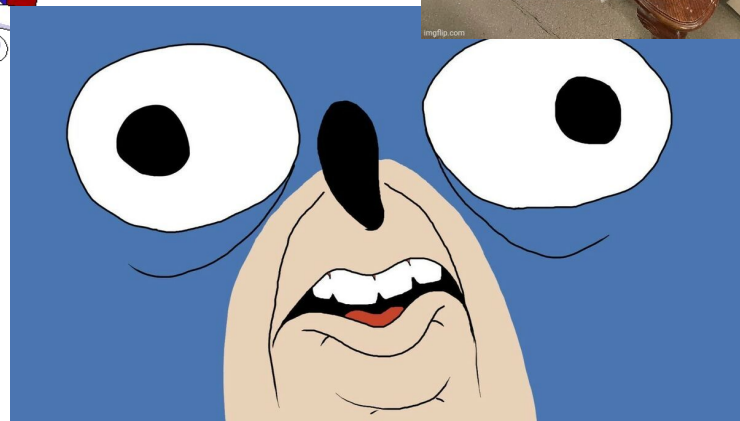
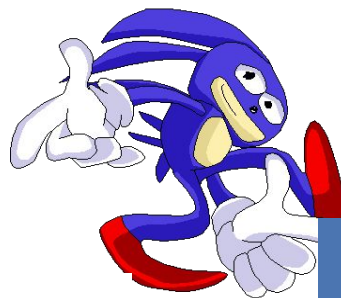
лучше сходится



Why so fast?

ТОЧКА

Мы умеем чистить **800 млн текстов** за **17 дней**
на текущем железе



1. Curated открытые датасеты **не такие уж и curated**

1. Curated открытые датасеты **не такие уж и curated**
2. Фильтрация данных значительно **ускоряет обучение** и **повышает конечную точность**

1. Curated открытые датасеты **не такие уж и curated**
2. Фильтрация данных значительно **ускоряет обучение** и **повышает конечную точность**
3. Old but gold: **trash in — trash out**



Елизавета Афанасьева

Senior Data Scientist в Точке

tg: @digitaljay

ML-каналчик

