DOI: 10.15514/ISPRAS-2025-37(3)-10

BERTScore для русского языка

[1] *Е.П. Бручес, ORCID: 0000-0002-1055-5339 <bruches@sibnn.ai>*
[1] *Д.Т. Батурова, ORCID: 0009-0000-1316-4838 <baturova@sibnn.ai>*
[2] *И.Ю. Бондаренко, ORCID: 0000-0002-0457-0698 <bondarenko@sibnn.ai>*

[1] *Сибирские Нейросети,*
Россия, 630090, г. Новосибирск, ул. Николаева, д. 12.

[2] *Новосибирский государственный университет,*
Россия, 630090, г. Новосибирск, ул. Пирогова, д. 2.

Аннотация. В работе представлено исследование по выбору наиболее релевантных векторных представлений для текстов на русском языке, которые используются в метрике BERTScore. Эта метрика используется для оценки качества сгенерированных текстов, которые могут быть получены в результате решения таких задач, как автоматическое реферирование текстов, машинный перевод и др.

Ключевые слова: метрика BERTScore; генерация текстов; автоматическая оценка.

Для цитирования: Бручес Е.П., Батурова Д.Т., Бондаренко И.Ю. BERTScore для русского языка. Труды ИСП РАН, том 37, вып. 3, 2025 г., стр. 147–158. DOI: 10.15514/ISPRAS–2025–37(3)–10.

Благодарности: Исследование было выполнено при поддержке программы "Приоритет 2030" национального проекта "Наука и университеты".

BERTScore for Russian

[1] *E.P. Bruches, ORCID: 0000-0002-1055-5339 <bruches@sibnn.ai>*
[1] *D.T. Baturova, ORCID: 0009-0000-1316-4838 <baturova@sibnn.ai>*
[2] *I.Y. Bondarenko, ORCID: 0000-0002-0457-0698 <bondarenko@sibnn.ai>*

[1] *Siberian Neuronets,*
12, Nikolaeva st., Novosibirsk, 630090, Russia.

[2] *Novosibirsk State University,*
2, Pirogova st., Novosibirsk, 630090, Russia.

Abstract. This paper presents a study on the selection of the most relevant vector representations for texts in Russian, which are used in the BERTScore metric. This metric is used to assess the quality of generated texts, which can be obtained as a result of solving tasks such as automatic text summarization, machine translation, etc.

Keywords: BERTScore metric; text generation; automatic evaluation.

For citation: Bruches E.P., Baturova D.T., Bondarenko I.Y. BERTScore for Russian. Trudy ISP RAN/Proc. ISP RAS, vol. 37, issue 3, 2025. pp. 147-158 (in Russian). DOI: 10.15514/ISPRAS-2025-37(3)-10.

Acknowledgments. This study was carried out with the support of the “Priority 2030” program of the national project “Science and Universities”.

1. Введение

Одними из популярных задач в области обработки естественных языков являются задачи генерации текстов: это может быть машинный перевод, суммаризация текстов и другие действия. Оценка качества решения данной задачи гораздо менее формализована, чем, например, задачи классификации, так как в данном случае следует в большей мере учитывать семантику текста, чем его формальные характеристики. Это обусловлено тем, что одна и та же мысль может быть по-разному выражена в текстовой форме, порядок слов может быть изменён и так далее.

С появлением BERT-моделей была предложена метрика BERTScore [1] для оценки качества сгенерированных текстов, которая основана на схожести эталонного текста и сгенерированного текста с использованием контекстных векторных представлений слов. Основная сложность использования этой метрики состоит в выборе наиболее релевантной модели для получения векторных представлений. Так, некоторые модели были обучены для определённых задач, либо конкретных предметных областей, поэтому их использование для других текстов может быть нерелевантным и давать менее точную оценку.

В статье [1] авторами проведено исследование, направленное на выявление корреляций между оценками BERTScore с использованием различных моделей и экспертными (человеческими) оценками. Таким образом, авторы получили статистические данные о зависимостях, что легло в основу рекомендаций об использовании той или иной модели в разных языках. К сожалению, это исследование было посвящено только английскому, китайскому и турецкому языкам, что не позволяет сделать выводы о том, какая модель наиболее подходит для оценки качества текстов на русском языке, который отличается от упомянутых языков флективностью и имеет свои грамматические и синтаксические особенности.

Настоящая работа посвящена исследованию корреляций между моделями из набора BERTScore и экспертной оценкой для выработки рекомендаций по использованию тех или иных моделей для оценки качества генерации текстов.

Работа состояла из следующих этапов:

1) Подготовка наборов данных, которые различаются между собой по типам задач, доменам и жанрам;
2) Генерация ответов моделей для данных задач;
3) Получение экспертной оценки для полученных ответов;
4) Расчет корреляций между значениями BERTScore (с векторными представлениями от разных моделей) и экспертной оценкой;
5) Выработка рекомендаций по использованию моделей для тех или иных задач.

2. Обзор литературы

Существует несколько групп метрик для оценки качества сгенерированных текстов: традиционные, статистические, нейросетевые и метрики, использующие языковые модели в качестве оценщиков.

Традиционные метрики включают в себя такие метрики как BLEU [2], ROUGE [3], METEOR [4] и некоторые другие.

- Метрика BLEU (BiLingual Evaluation Understudy) основана на подсчете слов и словосочетаний из сгенерированных текстов и эталонных текстов. Также она применяет штраф за краткость, чтобы избежать слишком высоких значений для кратких и неполных текстов.
- Метрика ROUGE (Recall-Oriented Understudy for Gisting Evaluation) похожа на метрику BLEU, но в её рамках считается не только метрика точности (precision), как в BLEU, но и метрика полноты (recall) и метрика F1, что позволяет обойтись без штрафа за краткость.
- Метрика METEOR (Metric for Evaluation of Translation with Explicit ORdering) также рассчитывает пересечения слов в двух текстах. Её отличие от предыдущих состоит в том, что она одновременно учитывает синонимы и однокоренные слова.
- Метрика TER (Translation Edit Rate) [5] основана на подсчете минимального числа правок, требуемых для приведения сгенерированного текста в полное соответствие наиболее близкому эталонному переводу.

Несмотря на относительную простоту в реализации этих метрик, главным их недостатком является отсутствие учёта семантики сгенерированного текста.

Статистические методы оценки качества основаны на сравнении распределений лексики двух текстов – сгенерированных автоматически и написанных людьми. Примерами таких метрик являются MAUVE [6] и GLTR [7].

Нейросетевые метрики оценки качества приобрели популярность в эпоху BERT моделей. Первая такая метрика – BertScore – была предложена в статье [1]. Данная метрика рассчитывает значение схожести (similarity score) для каждого слова в сгенерированном тексте с каждым словом в эталонном тексте. Но вместо полного посимвольного совпадения слов, используются контекстные векторные представления для каждого из них. Это позволяет учитывать семантику текстов даже при использовании разных слов, перефразировании и синтаксических различиях.

Похожая метрика MoverScore была предложена в статье [8]. Она также использует расчёт близости, используя векторные представления, но, кроме этого, также учитывает, насколько сгенерированный текст отклоняется от эталонного.

В работе [9] была предложена метрика BARTScore, основная идея которой заключается в расчёте вероятности появления каждого слова в сгенерированном тексте при условии определённого эталонного текста.

Использование языковых моделей в качестве оценщиков (LLM-as-a-Judge). С развитием больших языковых моделей, появилось такое актуальное направление исследований, как

использование моделей для автоматической оценки сгенерированных текстов. Примерами таких метрик являются TigerScore [10], INSTRUCTSCORE [11], Prometheus [12], Prometheus-2 [13] и другие. Такие метрики показали надёжность и достоверность при оценке сгенерированных текстов. Но этот подход имеет и некоторые недостатки, связанные с особенностями работы больших языковых моделей. Например, при таком использовании моделей остаётся проблема галлюцинирования [14]. Более того, у моделей могут оставаться "когнитивные" смещения, связанные с длиной контекста, с использованием определённых слов или фраз, порядком ответов и др. [15].

Подводя итог, отмечаем, что несмотря на возрастающую популярность подхода LLM-as-a-Judge, использование нейросетевых метрик остаётся востребованным, так как они вычислительно менее ресурсоёмки (по сравнению с использованием больших языковых моделей), но при этом учитывают семантику текстов в отличие от традиционных и статистических метрик.

3. Набор данных

Для определения наиболее релевантной модели, с помощью которой оценивается качество сгенерированного текста, мы собрали из открытых источников несколько наборов данных, которые различаются по задачам, предметным областям, длине текстов и другим характеристикам. Это позволит понять, есть ли необходимость в использовании разных моделей для разных задач. В данной работе мы использовали следующие наборы данных:

1) dialogsum-ru [16] – набор данных для задачи саммаризации диалогов. Исходный набор DialogSum [17] содержит 13 460 диалогов на английском языке на повседневные темы. Русскоязычный вариант набора был получен путём перевода текстов на русский язык с помощью Google Translate.
2) reviews-russian [18] – набор данных для суммаризации отзывов пользователей на отели и гостиницы. Содержит 92 текста.
3) ru-simple-sent-eval [19] – набор данных для задачи упрощения текстов на русском языке.
4) science-summarization-dataset [20] – набор данных для задачи суммаризации научных статей. Содержит 480 статей и их аннотаций из 8 различных научных областей: лингвистика, история, юриспруденция, медицина, компьютерные науки, экономика, химия.
5) telegram-financial-sentiment-summarization [21] – набор данных, который содержит тексты новостных сообщений из Телеграмма и их краткие содержания. Тексты в этом наборе, в основном, посвящены экономическим и политическим темам.
6) yandex-jobs [22] – набор данных, который содержит описания вакансий Яндекса. Задача заключается в генерации наиболее релевантного названия должности для каждой из вакансий.

В табл. 1 представлены следующие характеристики для каждого набора данных: средние длины входного и эталонного текстов, количество уникальных слов во входных и эталонных текстах. Из представленных описаний и статистик можно увидеть, что выбранные наборы данных, действительно, разнообразны по задачам, доменам и статистическим характеристикам. Таким образом, мы можем не только подобрать наиболее подходящую модель для оценки в среднем, но также и предложить наиболее релевантные модели для русского языка в зависимости от задачи и домена.

4. Модели для генерации ответов

Следующий этап данной задачи – сгенерировать ответы для описанных ранее задач. Для этого мы использовали две модели для русского языка: Gigachat Lite [23] и YandexGPT

Lite [24].

Для каждой из задач были вручную написаны подводки (prompts) для моделей, которые представлены в табл. 2. Для каждой модели была использована одна и та же подводка, без изменений.

Табл. 1. Характеристики наборов данных.
Table 1. Datasets characteristics.

Набор данных	Средняя длина входного текста	Средняя длина выходного текста	Количество уникальных слов во входных текстах	Количество уникальных слов в выходных текстах
dialogsum-ru	757	117	9 907	5 400
reviews-russian	1 390	448	5 096	2 115
ru-simple-sent-eval	137	91	15 300	20 694
science-summarization-dataset	20 155	843	112 380	13 469
telegram-financial-sentiment-summarization	313	169	26 818	20 426
yandex-jobs	925	38	8 364	504

Табл. 2. Подводки для моделей.
Table 2. Prompts for the models

Набор данных	Подводка для модели
dialogsum-ru	Сгенерируй краткий пересказ этого диалога в 1-2 предложениях:
reviews-russian	Сгенерируй краткий пересказ этого отзыва. Выделяй только важные факты. Отзыв:
ru-simple-sent-eval	Перепиши предложение так, чтобы его стало проще понимать:
science-summarization-dataset	Ниже приведена научная статья. Выдели главные факты и напиши краткое содержание этой статьи.
telegram-financial-sentiment-summarization	Напиши краткий пересказ этой новости. Выделяй только важные факты. Текст новости:
yandex-jobs	Напиши название должности для этой вакансии:

Из-за ограничений на генерацию текстов по неоднозначным темам обеих моделей (в разной степени) не для каждого текста удалось получить сгенерированный ответ. В табл. 3 представлено количество сгенерированных ответов каждой из моделей.

Таким образом, на данном этапе мы получили ответы от двух русскоязычных моделей для различных задач.

5. Модели для оценки ответов

В оригинальной статье BERTScore [1] расчёт корреляций происходил на основе экспертных

оценок: для используемых наборов данных эксперты вручную оценивали сгенерированные тексты. При том, что этот подход дает высокую точность оценки, это трудоёмкий процесс, который требует много времени и ресурсов. Ввиду этого, мы применили подход LLM-as-a-Judge, чтобы сгенерировать оценки с помощью языковой модели, не прибегая к экспертной ручной оценке.

Табл. 3. Количество сгенерированных каждой моделью текстов.
Table 3. The number of generated texts for the models.

Набор данных	GigaChat	YandexGPT
dialogsum-ru	267	1 494
reviews-russian	110	92
ru-simple-sent-eval	6 803	6 491
science-summarization-dataset	199	436
telegram-financial-sentiment-summarization	6 126	7 528
yandex-jobs	528	625

Стоит обратить внимание на проблему, которая возникает при таком использовании больших языковых моделей – это стремление моделей оценивать свои ответы выше, чем ответы других моделей [25, 26]. Поэтому в данной работе мы делали “перекрёстную” проверку – ответы YandexGPT оценивали с помощью модели GigaChat и наоборот, ответы GigaChat оценивали с помощью модели YandexGPT. Одним из основных ограничений такого подхода остаётся способность моделей галлюцинировать, то есть генерировать тексты, которые не имеют смысла или не соответствуют действительности, что может приводить к генерации моделями неверных оценок.

Для моделирования оценок от нескольких экспертов, каждая модель оценивала один и тот же текст тремя различными подводками. Как показано в статье [27], ансамбль подводок позволяет разнообразить оценки, уменьшить потенциальное смещение (то есть предотвращать генерацию одинаковых текстов для одной и той же подводки, что позволяет получать более разнообразные оценки) и увеличить надёжность таких оценок. Подводки содержали описание задачи, которую решали модели, эталонный и сгенерированный ответы, инструкцию оценить текст по 5-балльной шкале, где 1 – сгенерированный текст не соответствует эталонному, 5 – сгенерированный текст полностью соответствует эталонному.

Стоит отметить, что для получения более надёжных результатов, можно использовать не только набор из различных подводок, но и включить больше языковых моделей в задачу оценки сгенерированных текстов. Это позволит минимизировать влияние галлюцинаций моделей на итоговый результат.

Таким образом, в результате этого этапа мы собрали по 3 оценки (в некоторых случаях их получилось меньше из-за отказа моделей обрабатывать входной текст, цензурирования и других причин) для каждого сгенерированного текста.

6. Корреляции

В качестве коэффициента корреляции мы использовали тот, который был использован в оригинальной статье – коэффициент корреляции Пирсона.

Для каждого сгенерированного текста нами были подсчитаны значения метрики BERTScore с использованием всех доступных в инструменте моделей, которые поддерживают русский язык, а именно:

1) bert-base-multilingual-cased;
2) xlm-mlm-100-1280;
3) distilbert-base-multilingual-cased;

4) xlm-roberta-base;
5) xlm-roberta-large;
6) facebook/mbart-large-cc25;
7) facebook/mbart-large-50;
8) facebook/mbart-large-50-many-to-many-mmt;
9) google/mt5-small;
10) google/mt5-base;
11) google/mt5-large;
12) google/mt5-xl;
13) google/byt5-small;
14) google/byt5-base;
15) google/byt5-large;
16) microsoft/mdeberta-v3-base.

Для каждой модели использовались векторные представления, полученные на каждом слое. Таким образом, мы сможем подобрать не только наиболее релевантную модель, но и слой модели, векторные представления которого дают наиболее высокий коэффициент корреляции.

Кроме того, для каждой пары “эталонный текст – сгенерированный текст” мы посчитали стандартные метрики качества генерации: BLEU, ROUGE-1, ROUGE-2, ROUGE-L и MoverScore, чтобы сравнить корреляции BERTScore-метрик и статистических метрик с “экспертной” разметкой.

Также мы предоставляем значения корреляций для наиболее релевантных векторных представлений в среднем по всем наборам данных ("Лучший вектор" с указанием модели и слоя и "Лучшая модель" с указанием лучшей модели). Полученные корреляции представлены для моделей Gigachat и YandexGPT в табл. 4 и табл. 5 соответственно.

7. Анализ результатов

Из полученных результатов мы можем видеть, что семантические оценки качества, действительно, лучше коррелируют с экспертными, чем те, что основаны на формальных и статистических признаках (такие как BLEU и ROUGE-N). Кроме того, очевидно, что для разных доменов, жанров и задач наиболее релевантными оказываются векторные представления, полученные от различных моделей. Другим важным фактором выбора векторных представлений является сама модель, с помощью которых генерируются тексты.

Тем не менее, по результатам проведённых исследований показано, что в среднем наиболее релевантными векторными представлениями для всех доменов, задач и моделей являются те, которые получены из модели *google/byt5-base*. Эта модель является “token-free”, что означает, что она оперирует “сырым” текстом (байтами или знаками) вместо того, чтобы работать с отдельными словами, как большинство других языковых моделей. Такой подход даёт ряд преимуществ: возможность обрабатывать текст на любом языке без обучения, устойчивость к шумам и так далее. Возможно, именно этим обусловлена стабильность данной модели в среднем по всем наборам данных, которые были рассмотрены в данной работе. Кроме того, эта модель является компактной по числу параметров в сравнении с другими представленными моделями.

Также мы предоставляем результаты с наиболее релевантными векторными представлениями для каждой модели в среднем по всем наборам данных и моделям в табл. 6. Из приведённой таблицы видно, что, в среднем, наиболее релевантные векторные представления для оценки качества текстов на русском языке принадлежат моделям

google/byt5-base и *google/byt5-large*.

Табл. 4. Коэффициенты корреляции Пирсона для различных метрик (для текстов, сгенерированных GigaChat).
Table 4. Pearson correlation coefficients for various metrics (for texts generated by GigaChat).

Набор данных	Лучший вектор	BERTScore	BLE U	ROUGE- 1	ROUGE- 2	ROUGE- L	MoverScore
telegram-financial-sentiment-summarization	microsoft/mdeberta-v3-base_layer_ 7	**0.827**	0.423	0.645	0.497	0.607	0.575
ru-simple-senteval l	google/byt5-large_layer_31	**0.615**	0.096	0.281	0.149	0.25	0.323
science-summarization-dataset	facebook/mbart-large-50_layer_10	**0.749**	0.282	0.599	0.481	0.560	0.452
dialogsum-ru	facebook/mbart-large-cc25_layer_1 1	**0.447**	0.158	0.236	0.114	0.247	0.182
reviews-russian	facebook/mbart-large-5 0-many-to- many-mmt_ layer_6	**0.678**	0.178	0.346	0.206	0.346	0.545
yandex-j obs	google/byt5-base_layer_6	**0.433**	0.078	0.192	0.165	0.192	0.382
Лучший вектор	google/byt5-large_layer_29	**0.594**	0.191	0.379	0.257	0.359	0.476
Лучшая модель	google/byt5-large	**0.561**	0.191	0.379	0.257	0.359	0.476

8. Заключение

Автоматическая оценка качества сгенерированных текстов всё ещё остаётся сложной задачей. Несмотря на то, что за годы существования этого направления были предложены различные метрики, все они обладают своими недостатками. На данный момент, метрика BERTScore кажется наиболее оптимальной, так как, с одной стороны, учитывает семантические характеристики текстов, а с другой стороны, не требует такого большого количества ресурсов, как современные большие языковые модели.

Однако значения этой метрики очень зависят от выбранной модели, языка текстов и домена. В рамках данной работы мы провели исследование на выявление наиболее релевантной модели для оценки качества сгенерированных текстов для русского языка.

Ввиду того, что получение экспертных оценок для выявления корреляций является трудоёмким процессом, мы использовали подход LLM-as-a-Judge, где в качестве экспертов выступают большие языковые модели. Остаётся важным вопрос о границах применимости такого подхода – насколько оценки моделей, действительно, соотносятся с оценками от людей. Получение ответа на этот вопрос является одним из приоритетных направлений нашего дальнейшего исследования. Кроме того, в данной работе не были рассмотрены многие модели, обученные специально для русского языка. Этот аспект также является задачей для будущего исследования.

Табл. 5. Коэффициенты корреляции Пирсона для различных метрик (для текстов, сгенерированных YandexGPT).
Table 5. Pearson correlation coefficients for various metrics (for texts generated by YandexGPT).

Набор данных	Лучший вектор	BERTScore	BLE U	ROUGE- 1	ROUGE- 2	ROUGE- L	MoverScore
telegram-financial-sentiment-summarization	microsoft/mdeberta-v3-base_layer_8	**0.299**	0.157	0.238	0.197	0.239	0.233
ru-simple-senteval	xlm-roberta-large_layer_16	**0.224**	0.068	0.079	0.051	0.067	0.201
science-summarization-dataset	distilbert-base-multiling ual-cased_layer_5	**0.217**	0.017	0.116	0.134	0.122	0.115
dialogsum-ru	google/mt5-xl_layer_23	**0.291**	0.077	0.149	0.157	0.155	0.069
reviews-russian	google/mt5-xl_layer_23	**0.418**	0.138	0.234	0.176	0.213	0.289
yandex-j	google/byt5	**0.509**	0.169	0.303	0.214	0.301	0.416
obs	-base_layer_10						
Лучший вектор	google/byt5-base_layer_10	**0.284**	0.096	0.171	0.138	0.166	0.214
Лучшая модель	facebook/mbart-large-5 0-many-to- many-mmt	**0.249**	0.096	0.171	0.139	0.166	0.214

Табл. 6. Результаты для каждой из моделей.
Table 6. Results for each of the models.

Модель	Кол-во параметров	Слой	Pearson
google/byt5-base	528M	8	0.442
google/byt5-large	1.23B	29	0.442
facebook/mbart-large-50-many-to-many-mmt	611M	10	0.433
facebook/mbart-large-50	611M	10	0.430
google/mt5-xl	3.7B	23	0.421
microsoft/mdeberta-v3-base	280M	8	0.421
google/mt5-large	1.2B	22	0.413
facebook/mbart-large-cc2	610M	9	0.410
xlm-mlm-100-1280	570M	15	0.403
bert-base-multilingual-cased	179M	6	0.401
xlm-roberta-base	279M	6	0.397
xlm-roberta-large	561M	16	0.389
distilbert-base-multilingual-ca sed	135M	3	0.376
google/mt5-small	300M	3	0.371
google/mt5-base	580M	4	0.365
google/byt5-small	300M	1	0.341

Список литературы / References

[1]. Zhang T., Kishore V., Wu F., Weinberger K., Artzi Y. BERTScore: Evaluating Text Generation with BERT. International Conference on Learning Representations, 2020.

[2]. Papineni K., Roukos S., Ward T., Zhu W. Bleu: a Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[3]. Lin Ch. ROUGE: A Package for Automatic Evaluation of Summaries. Text Summarization Branches Out, 2004, pp. 74–81.

[4]. Banerjee S., Lavie Al. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 2005, pp. 65–72.

[5]. Snover M., Dorr B., Schwartz R., Micciulla L., Makhoul J. A Study of Translation Edit Rate with Targeted Human Annotation. Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, 2006, pp. 223–231.

[6]. Pillutla K., Swayamdipta S., Zellers R., Thickstun J., Welleck S., Choi Y., Harchaoui Z. MAUVE: measuring the gap between neural text and human text using divergence frontiers. Proceedings of the 35th International Conference on Neural Information Processing Systems, 2024, pp. 4816–4828.

[7]. Gehrmann S., Strobelt H., Rush Al. GLTR: Statistical Detection and Visualization of Generated Text. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2019, pp. 111–116.

[8]. Zhao W., Peyrard M., Liu F., Gao Y., Meyer Ch., Eger S. MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 563–578.

[9]. Yuan W., Neubig G., Liu P. BARTScore: Evaluating generated text as text generation. Thirty-Fifth Conference on Neural Information Processing Systems, 2021.

[10]. Jiang D., Li Y., Zhang G., Huang W., Lin B., Chen W. (2023) TIGERScore: Towards Building Explainable Metric for All Text Generation Tasks (online). Available at: https://arxiv.org/abs/2310.00752, accessed 28.10.2024.

[11]. Xu W., Wang D., Pan L., Song Zh., Freitag M., Wang W., Li L. INSTRUCTSCORE: Towards Explainable Text Generation Evaluation with Automatic Feedback. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 5967–5994.

[12]. Kim S., Shin J., Cho Y., Jang J., Longpre Sh., Lee H., Yun S., Shin S., Kim S., Thorne J., Seo M. (2023) Prometheus: Inducing Fine-grained Evaluation Capability in Language Models (online). Available at: https://arxiv.org/abs/2310.08491, 28.10.2024.

[13]. Kim S., Suk J., Longpre Sh., Lin B., Shin J., Welleck S., Neubig G., Lee M., Lee K., Seo M. (2024) Prometheus 2: An Open Source Language Model Specialized in Evaluating Other Language Models (online). Available at: https://arxiv.org/abs/2405.01535, accessed 28.10.2024.

[14]. Vu T., Krishna K., Alzubi S., Tar Ch., Faruqui M., Sung Y. (2024) Foundational Autoraters: Taming Large Language Models for Better Automatic Evaluation (online). Available at: https://arxiv.org/abs/2407.10817, accessed 28.10.2024.

[15]. Koo R., Lee M., Raheja V., Park J., Kim Z., Kang D. Benchmarking Cognitive Biases in Large Language Models as Evaluators. Findings of the Association for Computational Linguistics ACL 2024, 2024, pp. 517–545.

[16]. DIALOGSum Corpus, Available at: https://huggingface.co/datasets/d0rj/dialogsum-ru, accessed 11.02.2025.

[17]. Chen Y., Liu Y., Chen L., Zhang Y. DialogSum: A Real-Life Scenario Dialogue Summarization Dataset. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, pp. 5062–5074.

[18]. Reviews Russian, Available at: https://huggingface.co/datasets/trixdade/reviews_russian, accessed 11.02.2025.

[19]. Sakhovskiy A., Izhevskaia A., Pestova A., Tutubalina E., Malykh V., Smurov I., Artemova E. RuSimpleSentEval-2021 Shared Task: Evaluating Sentence Simplification for Russian. Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue” (2021), 2021, Ch. 161, pp. 607–617.

[20]. Tsanda Al., Bruches El. (2024) Russian-Language Multimodal Dataset for Automatic Summarization of Scientific Papers (online). Available at: https://arxiv.org/abs/2405.07886, accessed 28.10.2024.

[21]. Telegram Financial Sentiment Summarization dataset, Available at: https://huggingface.co/datasets/mxlcw/telegram-financial-sentiment-summarization, accessed 11.02.2025.
[22]. Yandex Jobs dataset, Available at: https://huggingface.co/datasets/Kirili4ik/yandex_jobs, accessed 11.02.2025.
[23]. GigaChat, Available at: https://giga.chat/, accessed 11.02.2025.
[24]. YandexGPT, Available at: https://ya.ru/ai/gpt-3, accessed 11.02.2025.
[25]. Bai Y., Ying J., Cao Y., Lv X., He Y., Wang X., Yu J., Zeng K., Xiao Y., Lyu H., Zhang J., Li J., Hou L. Benchmarking Foundation Models with Language-Model-as-an-Examiner. Advances in Neural Information Processing Systems, 2024.
[26]. Liu Y., Moosavi N., Lin Ch. LLMs as Narcissistic Evaluators: When Ego Inflates Evaluation Scores. In Findings of the Association for Computational Linguistics ACL 2024, 2024, pp. 12688–12701.
[27]. Shao W., Lei M., Hu Y., Gao P., Zhang K., Meng F., Xu P., Huang S., Li H., Qiao Y., Luo P. (2023) TinyLVLM-eHub: Towards Comprehensive and Efficient Evaluation for Large Vision-Language Models (online). Available at: https://arxiv.org/abs/2308.03729, accessed 28.10.2024.

Информация об авторах / Information about authors

Елена Павловна БРУЧЕС – кандидат технических наук, ведущий инженер-исследователь в компании “Сибирские Нейросети”. Сфера научных интересов: обработка естественных языков, извлечение информации, языковые модели, машинное обучение.

Elena Pavlovna BRUCHES – Cand. Sci. (Tech.) in Computer Sciences and a leading engineer-researcher in Siberian Neuronets. Research interests: natural language processing, information extraction, language models, machine learning.

Дари Тимуровна БАТУРОВА – бакалавр по направлению «Мехатроника и робототехника» Новосибирского государственного университета, разработчик-исследователь в компании «Сибирские нейросети». Сфера научных интересов: глубокое обучение, нейронные сети, обработка естественного языка, машинный перевод.

Dari Timurovna BATUROVA – Bachelor's degree in Mechatronics and Robotics at Novosibirsk State University, developer-researcher at Siberian Neural Networks company. Research interests: deep learning, neural networks, natural language processing, machine translation.

Иван Юрьевич БОНДАРЕНКО – научный сотрудник лаборатории прикладных цифровых технологий Новосибирского государственного университета, старший преподаватель кафедры фундаментальной и прикладной лингвистики Новосибирского государственного университета, сооснователь компании "Сибирские нейросети". Сфера научных интересов: машинное обучение, нейронные сети, компьютерная лингвистика, распознавание речи.

Ivan Yurevich BONDARENKO is a researcher in the Laboratory of Applied Digital Technologies and a senior lecturer in the Department of Fundamental and Applied Linguistics at Novosibirsk State University. He is also the co-founder of the Siberian Neural Networks company. Research interests: machine learning, neural networks, computational linguistics, speech recognition.