NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

as a manuscript

Vladislav Mikhailov

BENCHMARKING TRANSFORMER LANGUAGE MODELS ON NATURAL LANGUAGE UNDERSTANDING TASKS

PhD Dissertation Summary
for the purpose of obtaining academic degree
Doctor of Philosophy in Computer Science

Moscow — 2023

The PhD Dissertation was prepared at the National Research University Higher School of Economics.

Academic Supervisor: Ekaterina Artemova, Candidate of Sciences, National Research University Higher School of Economics, LMU Munich.

1 Introduction

Topic of the thesis

Natural language processing (NLP) is an interdisciplinary subfield of computational linguistics, computer science, and artificial intelligence aimed at the development of language technologies for performing tasks that involve the use of knowledge of the language, such as machine translation, question answering, information extraction, grammar error detection, and summarisation [31]. Large language models (LLMs) based on the Transformer architecture [117] have become an integral part of solutions for these tasks, leading to a paradigm shift in the area. The LLMs, also called the foundation models [15], are pretrained in a self-supervised manner at scale on a vast amount of text data and efficiently adapted to downstream tasks via finetuning [33; 64] and few-shot learning [17]. The rapid development and proliferation of the LLMs necessitate standardised methodologies for objectively evaluating their generalisation abilities across tasks, domains, and languages.

Benchmarking has found broad acceptance in computer science since the 1960s as a conventional approach to comparing systems with respect to specific criteria, such as performance, computational efficiency, security, and resilience [55; 60]. A benchmark represents a combination of one or more datasets associated with performance metrics and an aggregation procedure for summarising the results. More than 2,000 influential benchmarks[1] have been created by the NLP community to foster the development of general-purpose LLMs and address diverse aspects of the evaluation, including but not limited to general language understanding [119; 121], linguistic competence [125], cross-lingual generalisation [7], robustness to adversarial attacks [122], computational efficiency [143], and biases against disadvantaged social groups [74]. Most NLP benchmarks are gamified with public leaderboards, which enable a competitive evaluation of the LLMs against one another and human-level performance. Although benchmarking has become more application-oriented [62; 91], it suffers from low linguistic diversity [50] and the inappropriateness of the result aggregation procedures [27].

This dissertation is devoted to benchmarking Transformer LLMs on natural language understanding (NLU) tasks. We propose the first large-scale benchmarks for the Russian language that cover a broad scope of NLU tasks: machine reading comprehension (MRC), word sense disambiguation, coreference resolution, natural language inference, acceptability classification, authorship attribution, and artificial text detection. The latter is of particular interest to the fast-growing area of natural language generation (NLG) due to the growing risks of misusing the generative LLMs for malicious purposes [128]. Together with a benchmark for detecting machine-generated content, we develop a novel approach to this problem that relies on topological data analysis (TDA; [20]). Last, we introduce a framework for aggregating the performance results in multi-task benchmarks and multi-criteria evaluation protocols based on the

[1] paperswithcode.com/area/natural-language-processing. **Access date**: March 6, 2023.

social choice theory [6]. The aggregation procedures can be efficiently utilised to rank NLP systems in various evaluation scenarios while being more reliable and robust than the commonly used Pythagorean mean aggregation procedures.

Relevance of the work

Benchmarks for Russian. The advancement of machine learning (ML) technologies is inseparable from reliable evaluation. The NLP field predominantly focuses on English and exhibits skewed data and evaluation resource distribution for more than 7,000 languages [13; 90]. The data scarcity problem is addressed within the cross-lingual knowledge transfer paradigm, where the multilingual LLM is finetuned on the train set in a high-resource language – most often English – and evaluated on the test set in another language [92]. Even though this paradigm opens up new perspectives, it has several drawbacks. The transfer performance depends on the linguistic similarity between the source and target language and the amount of the target language's data in the model's pretraining corpus [59]. At the same time, languages typologically close to English are well-represented in multilingual benchmarks, such as XGLUE [63] and XTREME [46; 93], and the other cover a small fraction of tasks due to the lack of high-quality annotated data.

Recent research has adapted the benchmarking methodologies for English to develop large-scale NLU benchmarks for many typologically diverse languages, such as Polish [94], Korean [80], Basque [116], Arabic [101], Slovene [137], Chinese [133], Japanese [56], Persian [53], and Indonesian [130]. However, Russian is one of the languages that have received the least attention concerning standardised evaluation resources [50]. To this end, we present three novel NLU benchmarks for the Russian language:

1. Russian SuperGLUE [105] is a collection of nine Russian language understanding datasets created from scratch and designed analogically to the SuperGLUE benchmark [119]. The tasks include MRC, word sense disambiguation, coreference resolution, natural language inference, and a broad-coverage entailment diagnostic for a fine-grained model interpretation. The results of evaluating the Transformer-based LLMs for Russian at the time of release indicate that these models perform far below humans. Within three years, the newly developed LLMs have matched or surpassed the human performance on particular tasks, but remain inferior to humans by up to 4.9 of the overall score.
2. RuCoLA (Russian Corpus of Linguistic Acceptability; [70]) tests the linguistic competence of the LLMs with acceptability judgments, which reflect a sentence's well-formedness and naturalness from the perspective of native speakers [22]. RuCoLA consists of in-domain sentences manually collected from linguistic publications and out-of-domain sentences generated with nine downstream neural models. The out-of-domain set is developed to facilitate the practical use of acceptability judgments for improving Russian language generation. We empirically show that (i) the most widely used LLMs for Russian fall behind humans by

a large margin, especially when detecting morphological and semantic errors, and (ii) the cross-lingual knowledge transfer across Russian, English [126], and Italian [111] is hardly possible for the in-domain set. In contrast, the difference between monolingual and multilingual finetuning results for the out-of-domain set is less significant, meaning that the LLMs generalise well to judge the generated sentences.

3. RuATD (Russian Artificial Text Detection; [103]) is a multi-domain benchmark comprised of human-written and machine-generated texts. The neural texts are produced by 13 generative LLMs finetuned for text summarisation, paraphrase generation, text simplification, and machine translation. We also consider the back-translation and open-ended generation approaches. The RuATD benchmark has been organised as a shared task on (i) detection of neural texts, i.e., predicting whether a given text is natural or neural, and (ii) authorship attribution, i.e., identification of the author of a given text. The shared task has featured 38 submissions, with a performance gap of about 20% accuracy between the best-performing and least-performing ones for both task formulations. The evaluation results show that humans struggle to distinguish between the natural and neural texts while the detectors can achieve up to 83% accuracy.

Detection of neural texts. Disclaimer: The text in brown is generated with ChatGPT[2] to illustrate the need to develop generalisable artificial text detectors. The LLMs have become a powerful tool for generating text that closely resembles human language, but their misuse can have serious consequences. Misuse can lead to the amplification of biases present in the training data, the generation of misinformation, and privacy violations. Therefore, it is important to use these models responsibly, with careful consideration of the potential risks involved. Advancement of the LLMs enables new forms of misuse and stimulates the development of innovative approaches to mitigating risks of such misuse [15].

To address this line of research, we introduce a novel neural text detector based on TDA [57]. The TDA-based detector is a linear classifier trained on a concatenation of TDA features extracted from the Transformer's attention map represented as a weighted graph. The features include standard graph properties, descriptive characteristics of barcodes, and features based on the distance to attention patterns [25]. The experimental results show that the proposed detector outperforms count-based and BERT-based baselines [33] by up to 10% across three domains (Reddit, product reviews, and news) and is more generalisable to unseen GPT-2 models [86] than the baselines. The probing analysis of the features reveals their sensitivity to the surface and syntactic properties, which is analysed in greater detail in the follow-up work [21].

Aggregation procedures in benchmarking. The question of whether the mean aggregation procedure is suitable for ranking NLP systems in multi-task benchmarks remains a topic of

[2] openai.com/blog/chatgpt

ongoing debate. The mean aggregation simplifies the evaluation of the LLMs contrary to the considerable efforts of the community to keep benchmarking up-to-date. In particular, this procedure treats high-resource and low-resource languages equally and does not account for other criteria, such as task complexity and text domain [71; 127]. Moreover, the leading systems may outperform the others only on the outlier tasks, which leads to biased evaluation [1; 75].

Borrowing conventions from the social choice theory, we propose Vote'n'Rank [87], a framework that can be used to rank NLP systems in multi-task benchmarks and multi-criteria evaluation protocols. The framework includes eight aggregation procedures that account for the system rankings in each evaluation criterion and are suitable for aggregating heterogeneous performance measures. We conduct an empirical comparison of the Vote'n'Rank and Pythagorean mean procedures in four case studies: (i) re-ranking the GLUE, SuperGLUE, and VALUE [61] leaderboards, (ii) defining conditions that determine the system's top rank, (iii) assessing the procedures' robustness to missing task scores, and (iv) ranking NLP systems based on user preferences. The proposed aggregation procedures are more robust than the Pythagorean mean ones and provide interpretable decisions on the systems' rankings while accounting for missing performance scores and user preferences.

Research goal. The main goal of this work is to develop standardised evaluation resources and tools that (i) provide an exhaustive multi-domain comparison of existing and upcoming LLMs for Russian against the human-level performance, (ii) enable the inclusion of the Russian language into the cross-lingual research directions, and (iii) address the practical aspects of benchmarking, artificial text detection, and language generation evaluation.

2 Key results and conclusions

The **contributions** of this work can be summarized as follows:

1. We create the Russian SuperGLUE, RuCoLA, and RuATD benchmarks, which test the LLMs' generalisation abilities on 11 diverse NLU tasks across more than 15 text domains. We develop the methodologies for human evaluation, data collection, and data annotation accounting for specifics of the Russian language. Each benchmark hosts a public leaderboard for summarising the results of humans and state-of-the-art LLMs.
2. Together with the RuATD benchmark, we develop the TDA-based artificial text detector, which exploits geometrical properties underlying textual data and relies on structural differences in the topology of the Transformer LLMs' attention maps to distinguish between human-written and machine-generated texts.
3. We introduce Vote'n'Rank, a framework that includes eight aggregation procedures to rank LLMs in multi-task benchmarks and multi-criteria evaluation protocols under the

principles of the social choice theory. We provide recommendations for using the framework based on the procedures' properties and scenarios of the intended application.

4. We utilise the proposed evaluation resources and tools to conduct a detailed comparative analysis of more than 100 NLP systems, including count-based models, monolingual and multilingual Transformer LLMs, their ensembles, and other model configurations against the human-level performance in various experiment settings.

Theoretical and practical significance. We make application-oriented contributions based on the theoretical concepts of linguistics, TDA, and social choice theory. The following factors determine the *significance* of this thesis. We release the benchmarks, source code, leaderboards, human evaluation projects, and other materials under the Apache 2.0 license:

- Russian SuperGLUE (`GitHub repository`; `russiansuperglue.com`)
- RuCoLA (`GitHub repository`; `rucola-benchmark.com`)
- RuATD (`GitHub repository`)
    1. Detection of neural texts: `kaggle.com/competitions/ruatd-binary`
    2. Authorship Attribution: `kaggle.com/competitions/ruatd-authorship`
- The TDA-based detector (`GitHub repository`)
- Vote'n'Rank (`GitHub repository`)

The proposed benchmarks have become one of the standardised evaluation resources for Russian, with more than 2,000 private submissions received from the academic and industrial communities. In total, the public leaderboards rank more than 90 NLP systems against the human level, including widely used LLMs and their configurations, e.g., RuLeanALBERT[3], ruGPT-3[4], YaLM[5], FRED-T5[6], and ruRoBERTa[7]. The human evaluation projects can be re-used in many research directions, such as reproducibility of the human evaluation results [12], evaluating the effect of the project design on human performance [81], and analysing the performance differences between expert and non-expert annotators [51].

With Vote'n'Rank, researchers and practitioners can compare systems irrespective of the ML area. The framework allows the users to plug in their data and define their preferences in the evaluation. The evaluation resources and tools can also be used for educational purposes, e.g., practising the development of machine and deep learning models.

[3] `hf.co/yandex/RuLeanALBERT`

[4] `hf.co/ai-forever/rugpt3large_based_on_gpt2`

[5] `hf.co/yandex/yalm-100b`

[6] `hf.co/ai-forever/FRED-T5-1.7B`

[7] `hf.co/ai-forever/ruRoBERTa-large`

Last but not least, RuCoLA and RuATD promote the development of downstream and human-machine interaction models for evaluating the grammatical and semantic correctness in Russian language generation (e.g., ruRoBERTa-large-rucola[8]), detecting propaganda spread with bots, and warning users about potentially fake content on social media and news platforms.

Key aspects/ideas to be defended.

1. The Russian SuperGLUE, RuCoLA, and RuATD benchmarks as standardised evaluation resources for Russian.
2. An interpretable and robust ATD method based on TDA.
3. The Vote'n'Rank framework for ranking and determining single-winner NLP systems in multi-task benchmarks.
4. An empirical study of more than 100 LLMs and their configurations on NLU benchmarks.

Personal contribution. This thesis includes six publications, which result from the collaboration between authors of diverse backgrounds. In the first publication, "Russian SuperGLUE: A Russian Language Understanding Evaluation Benchmark" [105], the author of the thesis created RuCoS (Russian Reading Comprehension with Commonsense), the largest MRC dataset for Russian. The second publication, "Read and Reason with MuSeRC and RuCoS: Datasets for Machine Reading Comprehension for Russian" [38] describes in detail the approaches to the RuCoS collection and human evaluation through crowd-sourcing, which are developed solely by the author. The author also obtained the empirical results on the RuCoS dataset.

The author's contributions in the third paper, "RuCoLA: Russian Corpus of Linguistic Acceptability" [70], are (i) developing the high-level idea of the benchmark, (ii) collecting the in-domain sentences from the dataset on the Unified State Exam in the Russian language [104] and linguistic publications for a corpus-based description of Russian grammar, (iii) developing the methodologies for annotating the out-of-domain set and conducting estimates of the human performance jointly with Tatiana Shamardina, (iv) establishing the statistic and linguistic criteria for controlling the data quality, and (v) conducting the LLMs' performance and error analysis together with Max Ryabinin.

In the fourth paper, "Findings of the RuATD Shared Task 2022 on Artificial Text Detection in Russian" [103], the author (i) designed the benchmark and contributed as a Co-PI, (ii) aggregated the benchmark data collected by the co-authors, and (iii) developed the human evaluation project together with Tatiana Shamardina and Alena Fenogenova. The author's contributions in the fifth paper, "Artificial Text Detection via Examining the Topology of Attention Maps" [57] are (i) designing the experimental setup, (ii) conducting the attention head-wise probing, and (iii) analysing the results of each experiment.

[8] hf.co/RussianNLP/ruRoBERTa-large-rucola

In the sixth paper, "Vote'n'Rank: Revision of Benchmarking with Social Choice Theory" [87], the author (i) contributed as a Co-PI, (ii) designed the experimental setup, and (iii) conducted the first case study on re-interpreting NLP benchmarks with the assistance of Mark Rofin. Moreover, the author made the principal contributions to writing each paper.

Publications and probation of the work

* denotes equal contribution

First-tier publications

1. _Tatiana Shavrina, Alena Fenogenova, Anton Emelyanov, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev_. 2020. Russian SuperGLUE: A Russian Language Understanding Evaluation Benchmark. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4717–4726, Online. Association for Computational Linguistics. Conference rank: CORE A.
2. _Alena Fenogenova, Vladislav Mikhailov, and Denis Shevelev_. 2020. Read and Reason with MuSeRC and RuCoS: Datasets for Machine Reading Comprehension for Russian. In Proceedings of the 28th International Conference on Computational Linguistics (COLING), pages 6481–6497, Barcelona, Spain (Online). International Committee on Computational Linguistics. Conference rank: CORE A.
3. _Vladislav Mikhailov*, Tatiana Shamardina*, Max Ryabinin*, Alena Pestova, Ivan Smurov, and Ekaterina Artemova_. 2022. RuCoLA: Russian Corpus of Linguistic Acceptability. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 5207–5227, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. Conference rank: CORE A.
4. _Laida Kushnareva*, Daniil Cherniavskii*, Vladislav Mikhailov*, Ekaterina Artemova, Serguei Barannikov, Alexander Bernstein, Irina Piontkovskaya, Dmitri Piontkovski, and Evgeny Burnaev_. 2021. Artificial Text Detection via Examining the Topology of Attention Maps. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 635–649, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. Conference rank: CORE A.
5. _Mark Rofin*, Vladislav Mikhailov*, Mikhail Florinskiy*, Andrey Kravchenko, Elena Tutubalina, Tatiana Shavrina, Daniel Karabekyan, and Ekaterina Artemova_. 2023. Vote'n'Rank: Revision of Benchmarking with Social Choice Theory. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Dubrovnik, Croatia. Association for Computational Linguistics. Conference rank: CORE A.

Other publications

1. *Tatiana Shamardina*, Vladislav Mikhailov*, Daniil Cherniavskii, Alena Fenogenova, Marat Saidov, Anastasiya Valeeva, Tatiana Shavrina, Ivan Smurov, Elena Tutubalina, and Ekaterina Artemova*. 2022. Findings of the RuATD Shared Task 2022 on Artificial Text Detection in Russian. In Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2022". Indexed by Scopus.

Reports at conferences, workshops, and seminars

1. Russian SuperGLUE: A Russian Language Understanding Evaluation Benchmark. Online seminar at the Computational Pragmatics lab, HSE University.
2. Russian SuperGLUE: A Russian Language Understanding Evaluation Benchmark. EMNLP. November 17, 2020. Online presentation.
3. Read and Reason with MuSeRC and RuCoS: Datasets for Machine Reading Comprehension for Russian. COLING. December 8, 2020. Online presentation.
4. All Ways to Measure an Elephant: Russian SuperGLUE & RuSentEval. The International Symposium on the Application of Big Data Analysis for Trend Spotting. Session: Prospects for the Development of Applied Technologies for Big Data Processing and Natural Language Analysis. April 12, 2021. Online presentation.
5. Russian Commitment Bank: Machine Learning Lessons vs. Lessons of Linguistics – All not Learnt? Moscow HSE Pragmatics Workshop. September 30, 2021. Online presentation.
6. Artificial Text Detection via Examining the Topology of Attention Maps. EMNLP. Online and Punta Cana, Dominican Republic. November 7, 2021. Oral presentation.
7. Findings of the RuATD Shared Task 2022 on Artificial Text Detection in Russian. "Dialogue 2022". June 16, 2022. Online presentation.
8. RuCoLA: Russian Corpus of Linguistic Acceptability. EMNLP. December 9, 2022. Poster presentation.
9. Vote'n'Rank: Revision of Benchmarking with Social Choice Theory. EACL. May 2, 2023. Online presentation.

The author has organised the following conference events related to the thesis

1. *Tatiana Shavrina, Vladislav Mikhailov, Valentin Malykh, Ekaterina Artemova, Oleg Serikov, and Vitaly Protasov*. 2022. Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP. Association for Computational Linguistics (ACL), Dublin, Ireland. Conference rank: CORE A*.
2. *Adaku Uchendu, Vladislav Mikhailov, Jooyoung Lee, Saranya Venkatraman, Tatiana Shavrina, and Ekaterina Artemova*. 2022. Tutorial on Artificial Text Detection. The 15th International

Conference on Natural Language Generation (INLG), Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics. Conference rank: CORE B.

The author has also contributed to the following selected peer-reviewed publications

1. _Maria Tikhonova, Vladislav Mikhailov, Dina Pisarevskaya, Valentin Malykh, and Tatiana Shavrina_. 2022. Ad Astra or Astray: Exploring Linguistic Knowledge of Multilingual BERT Through NLI Task. In Natural Language Engineering, pages 1–30. Cambridge University Press. Journal Quartile: Q1.
2. _Daniil Cherniavskii*, Eduard Tulchinskii*, Vladislav Mikhailov*, Irina Proskurina*, Laida Kushnareva, Ekaterina Artemova, Serguei Barannikov, Irina Piontkovskaya, Dmitri Piontkovski, and Evgeny Burnaev_. 2022. Acceptability Judgements via Examining the Topology of Attention Maps. In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 88–107, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
3. _Ekaterina Taktasheva, Tatiana Shavrina, Alena Fenogenova, Denis Shevelev, Nadezhda Katricheva, Maria Tikhonova, Albina Akhmetgareeva, Oleg Zinkevich, Anastasiia Bashmakova, Svetlana Iordanskaia, Alena Spiridonova, Valentina Kurenshchikova, Ekaterina Artemova, and Vladislav Mikhailov_. 2022. TAPE: Assessing Few-shot Russian Language Understanding. In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 2472–2497, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
4. _Ekaterina Taktasheva, Vladislav Mikhailov, and Ekaterina Artemova_. 2021. Shaking Syntactic Trees on the Sesame Street: Multilingual Probing with Controllable Perturbations. In Proceedings of the 1st Workshop on Multilingual Representation Learning (MRL) at the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 191–210, Punta Cana, Dominican Republic. Association for Computational Linguistics.
5. _Vladislav Mikhailov, Oleg Serikov, and Ekaterina Artemova_. 2021. Morph Call: Probing Morphosyntactic Content of Multilingual Transformers. In Proceedings of the Third Workshop on Computational Typology and Multilingual NLP (SIGTYP) at the 2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), pages 97–121, Online. Association for Computational Linguistics.
6. _Vladislav Mikhailov, Ekaterina Taktasheva, Elina Sigdel, and Ekaterina Artemova_. 2021. RuSentEval: Linguistic Source, Encoder Force! In Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing (BSNLP) at the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL), pages 43–65, Kiyv, Ukraine. Association for Computational Linguistics.

Volume and structure of the work. This thesis contains an introduction, contents of publications, and a conclusion. The full volume of the thesis is 158 pages.

3 Content of the work

3.1 Russian SuperGLUE: A Russian Language Understanding Evaluation Benchmark

Dataset	$\vert$**Train**$\vert$	$\vert$**Dev**$\vert$	$\vert$**Test**$\vert$	**Task**	**Metrics**	**Domain**
DaNetQA	392	295	295	MRC	Acc.	Wikipedia
MuSeRC	500	100	322	MRC	F1$_a$/EM	news, fairy tales, academic texts, fiction, summaries of TV series and books
RuCoS	72k	4.3k	4.1k	MRC	F1/EM	news (Lenta, Deutsche Welle)
RUSSE	19.8k	8.5k	12.1k	WSD	Acc.	Wikipedia, RNC, dictionaries
PARus	400	100	500	NLI	Acc.	blogs, photography encyclopedia
RWSD	606	204	154	Coref.	Acc.	fiction
RCB	438	220	348	NLI	F1/Acc.	news, fiction
TERRa	2616	307	3198	NLI	Acc.	news, fiction
LiDiRus	✗	✗	1104	NLI	MCC	news, Wikipedia, Reddit, academic texts

Table 1: The tasks included in Russian SuperGLUE. LiDiRus is a diagnostic test set for the TERRa task. **Notations:** MRC=machine reading comprehension; WSD=word sense disambiguation; RNC=Russian National Corpus; Coref.=coreference resolution; and NLI=natural language inference. **Metrics:** F1=F1-score; F1$_a$=macro-average F1 [52]; Acc.=accuracy; EM=exact match; MCC=Matthews Correlation Coefficient [67].

Russian SuperGLUE shares the same motivation as the GLUE [121] and SuperGLUE [119] benchmarks: to introduce a standardised methodology for measuring the advancement of language understanding technologies for Russian. Russian SuperGLUE is a suite of eight NLU tasks that cover various text domains, dataset sizes, and task formulations (see Table 1). We provide an online platform and a public leaderboard for model evaluation, comparison, and analysis based on private test sets.

3.1.1 Method

DaNetQA (A Yes/No Question Answering Dataset) is a MRC task formulated as a binary classification problem in which the model has to answer a yes/no natural question to a given Wikipedia passage.

- **Text**: *"В период с 1969 по 1972 год по программе <Аполлон> было выполнено 6 полётов с посадкой на Луне. Всего на Луне высаживались 12 астронавтов США."*
- **Question:** *"Был ли человек на луне?"*
- **Answer:** `True`

MuSeRC (Russian Multi-Sentence Reading Comprehension) is a MRC task that focuses on multi-sentence inference and encompasses a range of question types that necessitate reasoning abilities.

- **Text:** *"(1) Мужская сборная команда Норвегии по биатлону в рамках этапа Кубка мира в немецком Оберхофе выиграла эстафетную гонку. (2) Вторыми стали французы, а бронзу получила немецкая команда. (3) Российские биатлонисты не смогли побороться даже за четвертое место, отстав от норвежцев более чем на две минуты. (4) Это худший результат сборной России в текущем сезоне. (5) Четвёртыми в Оберхофе стали австрийцы. (6) В составе сборной Норвегии на четвёртый этап вышел легендарный Уле-Эйнар Бьорндален. (7) Впрочем, Норвегия с самого начала гонки была в числе лидеров, успешно проведя все четыре этапа. <...> (11) Напомним, что днем ранее российские биатлонистки выиграли свою эстафету. (12) В составе сборной России выступали Анна Богалий-Титовец, Анна Булыгина, Ольга Медведцева и Светлана Слепцова. (13) Они опередили своих основных соперниц - немок - всего на 0,3 секунды."*
- **Question:** *"На сколько секунд женская команда опередила своих соперниц?"*
- **Answers:**

  ☑ *"Всего на 0,3 секунды."*
  ☑ *"На 0,3 секунды."*
  □ *"На секунду."*
  □ *"На 0.5 секунд."*

RuCoS (Russian Reading Comprehension with Commonsense Reasoning) is an MRC task requiring commonsense reasoning and world knowledge. Each dataset sample consists of a passage from a news article, a cloze-style query, and answer options. The task is to fill in the placeholder in the query by selecting one or more referents of the answer entity in the passage.

- **Passage:** *"Мать двух мальчиков, брошенных отцом в московском аэропорту Шереметьево, забрала их. Об этом сообщили ТАСС в пресс-службе министерства образования и науки Хабаровского края. Сейчас младший ребенок посещает детский сад, а старший ходит в школу. В учебных заведениях с ними по необходимости работают штатные психологи. Также министерство социальной защиты населения рассматривает вопрос о бесплатном оздоровлении детей в летнее время. Через несколько дней после того, как Виктор Гаврилов бросил своих детей в аэропорту, он явился с повинной к следователям в городе Батайске Ростовской области.*

  - *Бросившего детей в Шереметьево отца задержали за насилие над женой*
  - *Россиянина заподозрили в истязании брошенных в Шереметьево детей*
  - *Оставивший двоих детей в Шереметьево россиянин сам пришел к следователям"*

- **Question:** *"26 января ___ бросил сыновей в возрасте пяти и семи лет в Шереметьево."*
- **Answer:** *"Виктор Гаврилов"*

RUSSE (Russian Words in Context) is a word sense disambiguation task framed as a binary classification problem. The task is to identify whether a given polysemous word is used with the same sense in a pair of sentences.

- **Sentence 1:** “*Бурые ковровые дорожки заглушали шаги.*”
- **Sentence 2:** “*Приятели решили выпить на дорожку в местном баре.*”
- **Word:** “*дорожка*”
- **Answer:** `False`

PaRus (Choice of Plausible Alternatives for Russian) is causal reasoning task framed as a binary classification problem. The model is given a premise sentence and has to identify which of the two given alternatives represent either the cause or effect.

- **Premise:** “*Гости вечеринки прятались за диваном.*”
- **Choice 1:** “*Это была вечеринка-сюрприз.*”
- **Choice 2:** “*Это был день рождения.*”
- **Question:** “*Причина*”
- **Answer:** “*Это была вечеринка-сюрприз.*”

RWSD (Russian Winograd Schema Challenge) is a coreference resolution task that involves a sentence with a pronoun and a list of noun phrases. The model must identify the correct pronoun referent among the provided options.

- **Text:** “*Кубок не помещается в коричневый чемодан, потому что он слишком большой.*”
- **Pronoun/Noun phrase:** “*он слишком большой*”
- **Referent:** “*чемодан*”
- **Answer:** `False`

RCB (Russian Commitment Bank) is a textual entailment task cast as a three-class classification problem. Each dataset sample includes a premise with a clause-embedding predicate under an entailment cancelling operator and a hypothesis. The model is required to predict whether the hypothesis entails, contradicts, or is neutral to the premise.

- **Premise:** “*Сумма ущерба составила одну тысячу рублей. Уточняется, что на место происшествия выехала следственная группа, которая установила личность злоумышленника. Им оказался местный житель, ранее судимый за подобное правонарушение.*”
- **Hypothesis:** “*Ранее местный житель совершал подобное правонарушение.*”
- **Predicate:** “*судить*“
- **Answer:** `entailment`

Coarse-Grained Categories	**Fine-Grained Categories**
Lexical Semantics	Lexical Entailment, Morphological Negation, Factivity, Symmetry/Collectivity, Redundancy, Named Entities, Quantifiers
Predicate-Argument Structure	Core Arguments, Prepositional Phrases, Ellipsis/Implicits, Anaphora/Coreference, Active/Passive, Nominalization, Genitives/Partitives, Datives, Relative Clauses, Coordination Scope, Intersectivity, Restrictivity
Logic	Negation, Double Negation, Intervals/Numbers, Conjunction, Disjunction, Conditionals, Universal, Existential, Temporal, Upward Monotone, Downward Monotone, Non-Monotone
Knowledge	Common Sense, World Knowledge

Table 2: The linguistic phenomena in LiDiRus.

TERRa (Textual Entailment Recognition for Russian) is a binary classification problem that requires recognising whether the meaning of one text can be entailed from another in a given pair of texts.

- **Premise:** *"Автор поста написал в комментарии, что прорвалась канализация."*
- **Hypothesis:** *"Автор поста написал про канализацию."*
- **Answer:** `entailment`

LiDiRus (Linguistic Diagnostic for Russian) is a broad-coverage entailment diagnostic test set for the TERRa task, which covers a wide range of phenomena for a fine-grained model interpretation (see Table 2). LiDiRus aims to analyse the relationship between the model predictions and phenomena through correlation analysis.

- **Sentence 1:** *"Мы построили наше общество на неэкологичной энергии."*
- **Sentence 2:** *"Мы построили наше общество на экологичной энергии."*
- **Answer:** `not entailment`
- **Lexical Semantics:** `morphological negation`
- **Logic:** `negation`

3.1.2 Empirical evaluation

Baselines. We empirically evaluate count-based baselines and BERT-based LLMs for Russian. TF-IDF is a Logistic Regression classifier over TF-IDF features [96] computed on a subset of 20k Russian and English Wikipedia articles. mBERT [33] is a multilingual BERT pretrained on monolingual Wikipedia corpora in 104 languages. ruBERT-base [136] is a Russian BERT

Model	Overall	LiDiRus MCC	RCB F1/Acc.	PARus Acc.	MuSeRC $F1_a$/EM	TERRa Acc.	RUSSE Acc.	RWSD Acc.	DaNetQA Acc.	RuCoS F1/EM
TF-IDF	43.4	5.9	30.1/44.1	48.6	58.7/24.2	47.1	66.0	66.2	62.1	25.6/25.1
ruBERT	54.6	18.6	43.2/46.8	61.0	65.6/25.6	63.9	**89.4**	67.5	74.9	25.5/25.1
mBERT	54.2	15.7	38.3/42.9	58.8	62.6/25.3	62.0	84.0	67.5	79.0	37.1/36.7
Human	**80.2**	**62.6**	**68.0/70.2**	**98.2**	**80.6/42.0**	**92.0**	74.7	**84.0**	**87.9**	**93.0/92.4**

Table 3: Baseline performance on the Russian SuperGLUE private test sets and diagnostics. **Metrics**: F1=F1-score; $F1_a$=macro-average F1 [52]; Acc.=accuracy; EM=exact match; MCC=Matthews Correlation Coefficient [67]. All values are scaled by 100. The **Overall** column is the mean average benchmark score. The scores in bold reflect the best performance on dataset.

pretrained on Russian news and Wikipedia, with the weights initialised from the mBERT model. The BERT-based models are finetuned on the corresponding task. We also conduct estimates of the human performance on each dataset using Toloka[9], a crowd-sourcing platform for data labelling. Annotation instructions and examples of the web interface for each task are provided in the GitHub repository[10].

Key results. The results are presented in Table 3. Overall, the BERT-based LLMs significantly underperform humans on most tasks. However, the LLMs exceed the human level on the RUSSE task by up to a 14.7% accuracy score. Comparing the LLMs with one another, we observe that ruBERT performs slightly better than mBERT, especially on the NLI tasks (RCB, TERRa) and MRC tasks (RuCoS, DaNetQA).

3.1.3 Retrospective

Since its release, Russian SuperGLUE has undergone community validation and methodological improvements [39]: (i) enriching the RUSSE test set with 6.7k samples and improving the human performance by a 6% accuracy score, (ii) extending the MuSeRC train set with 300 samples, (iii) doubling the size of the RuCoS development and test sets, (iv) increasing the size of the DaNetQA dataset up to 1750, 821, and 805 samples in the train, development, and test sets, and (v) cleaning typos and inaccuracies and improving the annotation consistency through manual development of the MuSeRC and RuCoS datasets.

Within three years, the overall performance gap between humans and the leading LLMs has been narrowed from 25.8 to 4.9. The progress has been achieved due to the advances in language modelling and the development of novel Russian LLMs, such as RuLeanALBERT, ruGPT-3, YaLM, FRED-T5, and ruRoBERTa. Russian SuperGLUE has received over 2,000 private submissions from the academic and industrial communities and ranks 28 NLP systems on the

[9] toloka.ai

[10] github.com/RussianNLP/RussianSuperGLUE/HumanBenchmark

Rank	Model	Overall	LiDiRus MCC	RCB F1/Acc.	PARus Acc.	MuSeRC $F1_a$/EM	TERRa Acc.	RUSSE Acc.	RWSD Acc.	DaNetQA Acc.	RuCoS F1/EM
1	Human	**81.1**	**62.6**	**68.0**/**70.2**	**98.2**	80.6/42.0	**92.0**	80.5	**84.0**	**91.5**	**93.0**/89.0
2	FRED-T5 1.7B FT	76.2	49.7	49.7/54.1	84.2	91.6/77.3	87.1	**82.3**	66.9	88.9	90.0/90.2
3	Golden Transformer v2.0	75.5	51.5	38.4/53.4	90.6	**93.6**/**80.4**	87.7	68.7	64.3	91.1	92.0/**92.4**
4	YaLM p-tune	71.1	36.4	35.7/47.9	83.4	89.2/70.7	84.1	71.0	66.9	85.0	92.0/91.6
5	FRED-T5 large FT	70.6	38.9	45.6/54.6	77.6	88.7/67.8	80.1	77.5	66.9	79.9	87.0/86.3
6	RuLeanALBERT	69.8	40.3	36.1/41.3	79.6	87.4/65.4	81.2	78.9	66.9	76.0	90.0/90.2
7	FRED-T5 1.7B encoder FT	69.4	42.1	31.1/44.1	80.6	88.2/66.6	83.1	72.3	66.9	73.5	91.0/91.1
8	ruT5-large FT	68.6	32.0	45.0/53.2	76.4	85.5/60.8	77.5	77.3	66.9	79.0	86.0/85.9
9	ruRoBERTa-large FT	68.4	34.3	35.7/51.8	72.2	86.1/63.0	80.1	74.8	66.9	82.0	87.0/86.7

Table 4: Top-nine positions on the Russian SuperGLUE leaderboard. **Metrics**: F1=F1-score; $F1_a$=macro-average F1 [52]; Acc.=accuracy; EM=exact match; MCC=Matthews Correlation Coefficient [67]. FT stands for finetuning. All values are scaled by 100. The **Overall** column is the mean average benchmark score. The scores in bold reflect the best performance on dataset.

public leaderboard (see Table 4). Different approaches and LLMs' configurations have been evaluated on the benchmark, including standard finetuning of the encoder (ruRoBERTa-large, RuLeanALBERT) and encoder-decoder (ruT5-large, FRED-T5) LLMs, prompt-tuning (YaLM), and ensembles (Golden Transformer v2.0). Simple baselines, such as TF-IDF, random guessing, and majority classifier, take the last three positions. While the systems match or outperform humans on the word sense disambiguation (RWSD) and MRC tasks (MuSeRC, DaNetQA, and RuCoS), there is still room for improving the LLMs' generalisation to the Winograd Schema Challenge (RWSD), NLI (RCB, TERRa), and causal reasoning tasks (PARus).

3.2 Read and Reason with MuSeRC and RuCoS: Datasets for Machine Reading Comprehension for Russian.

MRC is one of the central NLU tasks with wide real-world applications that incorporates general language understanding, world knowledge, and logical reasoning to answer a question. Various task formulations have been proposed [139], including cloze-style (filling in the placeholders), multiple choice (selecting one answer from given options), span prediction (extracting a text segment), and free-form answer (answering in any free-text form). However, as in other NLP areas, this research field focuses on English [89].

At the time of the Russian SuperGLUE release, the task of MRC for Russian was primarily explored in the context of span prediction in the monolingual and cross-lingual scenarios [35; 7; 24]. In response, we propose MuSeRC and RuCoS, two Russian MRC datasets requiring multi-sentence reasoning and commonsense knowledge. MuSeRC follows the design of MultiRC [52], while RuCoS is modelled after ReCoRD [123]. This section (i) details methodologies for the RuCoS collection and human evaluation, (ii) provides a comparative analysis of the ReCoRD and RuCoS general statistics, and (iii) presents the results of evaluating three BERT-based LLMs and human annotators on RuCoS and MuSeRC.

3.2.1 Method

Data collection. We describe the methodology for creating the RuCoS dataset, consisting of passages with titles of related news articles, cloze-style queries, and answer options. The methodology includes five main steps: (1) collecting Lenta.ru[11] and Deutsche Welle[12] news articles, (2) generating <*passage, cloze-style query, answers*> triples, (3) filtering out passages with low-frequency words, (4) discarding samples that out-of-the-box MRC models can answer, and (5) filtering out samples ambiguous to human annotators.

1. **Lenta.ru and Deutsche Welle news article curation.** We parse news articles from Lenta.ru and Deutsche Welle and extract named entities (NEs) in the articles with a BERT-based named entity recognition model using the DeepPavlov library [19].
2. **Passage-query-answers generation.** We generate dataset samples in the form of <*passage, cloze-style query, answers*> triples. Each *passage* consists of the first few paragraphs of a news article and three titles of related news articles ranked by cosine similarity between the TF-IDF representations of the passage and titles. The titles provide additional context or complementary summary points. Each *query* is a sentence that follows the passage, contains at least one NE mentioned in the passage, and satisfies the criteria defined in [140]. We replace only one NE in the selected sentence with a placeholder to generate the cloze-style query. The *answer* options are the extracted NEs in the *passage*.
3. **Frequency Filtering.** We compute the token frequency in each paragraph as the number of frequently used tokens (i.e., the number of instances per million in the Russian National Corpus (RNC)[13] is higher than one) divided by the number of tokens in a paragraph. We keep triples that contain passages with more than 70% high-frequency tokens.
4. **Machine filtering.** We filter out triples correctly answered by at least one of the two extractive MRC models for Russian available via the DeepPavlov library: ruBERT and R-NET [123]. At this step, we randomly split the triples into train, dev, and private test sets, with the news source balanced.
5. **Human filtering.** We validate the resulting dev and test triples with the help of crowdsourcing workers on Toloka (see § A.1 in [38] for an example of the annotation instruction). The project includes an unpaid training phase with explanations, control tasks for tracking annotation quality, and the main annotation phase. Before starting, the worker is given detailed instructions describing the task and showing annotation examples. The instruction is available anytime during the training and main annotation phases.

[11] lenta.ru

[12] www.dw.com/ru/

[13] ruscorpora.ru/new/en

Parameter	ReCoRD				RuCoS			
	Train	Dev	Test	Overall	Train	Dev	Test	Overall
Num. of samples	65,709	7,481	7,484	80,674	72,193	7,577	7,257	87,027
Num. of queries	100,730	10,000	10,000	120,730	72,193	7,577	7,257	87,027
Num. of unique queries	99,713	9,977	9,968	80,179	72,193	7,577	7,257	87,027
Num. of unique passages	65,258	7,133	7,279	79,670	72,193	7,577	7,257	87,027
|Query vocabulary|	119,069	30,844	31,028	134,397	109,899	30,203	27,813	120,410
|Passage vocabulary|	352,491	93,171	94,386	395,356	279,333	90,699	83,237	303,647
Num. of tokens per query	21.3	22.1	22.2	21.4	22.2	22.1	21.6	22.2
Num. of tokens per passage	169.5	168.6	168.1	169.3	146.6	146.2	142.5	146.2
Num. of NEs per passage	17.2	17.3	17.2	17.2	12.7	14.3	13.3	12.9
NE frequency	7.1	4.4	4.3	7.5	8.9	5.0	5.3	9.6
Answer frequency	6.8	4.7	✗	6.5	10.2	4.1	✗	10.2
% high-frequency tokens per query	✗	✗	✗	✗	86.0	85.0	86.0	86.0
% high-frequency tokens per passage	✗	✗	✗	✗	82.0	81.0	82.0	82.0

Table 5: General statistics of the ReCoRD and RuCoS datasets.

Access to the project is granted to workers who (i) have a user rating of more than 60% and (ii) complete the training phase by labelling at least 7 out of 10 samples correctly. The annotation task is to (i) validate coherence between the passage and the query, (ii) report if the answer is not obvious or ambiguous, (iii) select all possible answers, and (iv) report any inconsistency and errors, e.g., an incomplete entity markup or misspellings. We keep submissions with more than 30 seconds of response time and discard votes from workers whose quality on the control tasks is higher than 50%. We collect the majority vote labels via a dynamic overlap[14] from three to five remaining workers. Two authors of the paper manually validated each submission and corrected all reported drawbacks.

General statistics. We use spaCy for English[15] and Russian[16] to compute the statistics of the ReCoRD and RuCoS datasets presented in Table 5. The distribution of samples by the news source is 44%/56% in ReCoRD (CNN/Daily Mail News) and 67%/33% in RuCoS (Lenta.ru/Deutsche Welle). RuCoS is larger than ReCoRD by 6.3k samples. Unlike ReCoRD, RuCoS has unique passages and queries in each dataset sample, meaning there is only one query for each passage. We observe that (i) passages in RuCoS are shorter, (ii) queries in RuCoS contain fewer NEs, and (iii) ReCoRD tends to be more diverse regarding the entity and answer vocabularies. This can be attributed to the language peculiarities, specifics of the data sources, and topic distribution. At the same time, RuCoS requires an understanding of rich inflectional morphology and high lexical variability in Russian.

[14] toloka.ai/docs/dynamic-overlap

[15] github.com/explosion/spaCy

[16] github.com/aatimofeev/spacy_russian_tokenizer

Model	MuSeRC $F1_a$/EM	RuCoS F1/EM
TF-IDF	58.9/24.4	25.6/25.1
mBERT	66.8/33.6	30.6/29.6
ruBERT-conv	71.7/32.9	26.4/25.9
ruBERT-base	71.7/33.6	34.4/33.9
Human	**80.6/42.0**	**93.0/92.4**

Table 6: Baseline evaluation on the RuCoS and MuSeRC datasets. **Metrics**: F1=F1-score; $F1_a$=macro-average F1 [52]; EM=exact match.

3.2.2 Empirical evaluation

Baselines. We experimentally evaluate count-based models and BERT-based LLMs for the Russian language: TF-IDF, ruBERT-base, mBERT, and ruBERT-conv[17]. Here, the TF-IDF approach includes (i) building the term vocabulary on the corresponding training set, (ii) replacing the query with each answer option (RuCoS) or concatenating the passage with each answer option (MuSeRC), (iii) computing the cosine similarity between the TF-IDF vectors of the passage and resulting query (RuCoS) or the concatenation and the question (MuSeRC), and (iv) returning the answer option that maximises the similarity. We evaluate the human performance on Toloka, where the workers are required to (i) read the passage and the cloze-style query with a placeholder, (ii) select all possible answers that best fit the placeholder, and (3) report ambiguous samples and errors. The annotation instructions and examples of the web interface for MuSeRC and ReCoRD are publicly available[18].

Metrics. The evaluation design follows the works by [123; 52]. Exact match (EM) measures the percentage of predictions that match all true answer options (MuSeRC) or any of the true answer options (RuCoS). Macro-average F1 ($F1_a$) is a harmonic mean of the precision and recall scores computed over binary decisions for each answer option. F1-score (F1) measures the overlap between the prediction and true answer options treated as bag-of-words. We compute the maximum F1 score for all reference answers per query and then average it across all queries.

Key results. Table 6 presents the evaluation results. The monolingual LLMs perform best on the MuSeRC dataset, while mBERT outperforms ruBERT-conv on the RuCoS dataset. ruBERT-base receives the best performance among the models on both tasks. We also find a significant difference between the human and baseline results, specifically on RuCoS.

[17] hf.co/DeepPavlov/rubert-base-cased-conversational

[18] github.com/RussianNLP/RussianSuperGLUE/HumanBenchmark

3.2.3 Retrospective

The empirical evaluation results demonstrate that the most widely-used Russian LLMs when writing the publication are significantly inferior to humans by up to 58.6 F1-score and 58.5 EM score. Nowadays, the MuSeRC and RuCoS datasets are *saturated*, meaning that the LLMs match or outperform humans. The best-performing models (see Table 4) rely on finetuning (FRED-T5, RuLeanALBERT), prompt-tuning (YaLM), and model ensembles (Golden Transformer v2.0). However, the proposed datasets have contributed to increasing the inclusivity of the Russian language, making it the third best-resourced language in the context of question answering and MRC problems [89].

3.3 RuCoLA: Russian Corpus of Linguistic Acceptability

The question of whether neural LLMs acquire grammatical knowledge central to human linguistic competence has been addressed with *acceptability judgments*, which reflect the degree to which a sentence is well-formed and natural from the perspective of native speakers [22]. Acceptability judgments have formed an empirical foundation in generative linguistics for evaluating humans' grammatical knowledge and language acquisition [99]. The NLP field has applied the conventions from linguistic theory to test model robustness [135], interpret the performance of downstream models [18], and evaluate the grammatical correctness in language generation [10; 11]. With the release of CoLA (Corpus of Linguistic Acceptability; [126]) included in the GLUE benchmark, the community has dedicated significant effort to creating similar linguistic acceptability resources in multiple languages, except for Russian [111; 42; 118; 132].

RuCoLA is the first large-scale acceptability classification benchmark in Russian, which combines in-domain sentences manually collected from linguistic literature and out-of-domain sentences generated with nine machine translation and paraphrase generation models. The motivation behind the out-of-domain set is to enable the application of acceptability judgments to enhance language generation in practical scenarios. RuCoLA provides a website and a public leaderboard for testing the linguistic competence of modern and upcoming LLMs for the Russian language.

3.3.1 Method

The task is framed as a sentence-level binary classification problem, where the model is required to predict whether the sentence is acceptable.

- **Sentence:** "*Иван прилёг, чтобы он отдохнул.*"
- **Answer:** `False` *(Неприемлемое предложение)*
- **Category:** *Синтаксис*
- **Source:** [109]

Source	Size	%	Content
rusgram	563	49.7	Corpus grammar
[109]	1,335	73.9	General syntax
[65]	193	75.6	Syntactic structures
[72]	54	57.4	Generative grammar
[78]	1,308	84.3	Semantics of tense
[77]	1,374	90.8	Lexical semantics
[79]	1,462	89.5	Aspects of negation
[102]	2,104	80.8	Semantics
[104]	1,444	36.6	Grammar exam tasks
In-domain	**9,837**	**74.5**	
Machine translation	1,286	72.8	English translations
Paraphrase generation	2,322	59.9	Automatic paraphrases
Out-of-domain	**3,608**	**64.6**	
Total	**13,445**	**71.8**	

Table 7: RuCoLA statistics by source. **Notations:** %=Percentage of acceptable sentences. rusgram is a collection of materials written by linguists for a corpus-based description of Russian grammar (available at: rusgram.ru).

Data collection. RuCoLA is composed of in-domain and out-of-domain sets (see Table 7). The *in-domain set* is created through a manual collection of sentences and the corresponding authors' acceptability judgments from linguistic textbooks, academic publications, and methodological materials on Russian grammar. The *out-of-domain set* consists of sentences produced by nine open-source machine translation (MT) and paraphrase generation (PG) models via the EasyNMT[19] and russian-paraphrasers [37] libraries on subsets of four datasets from different domains: Tatoeba [8], WikiMatrix [100], TED [85], and Yandex Parallel Corpus [5]. The MT models are OPUS-MT [110], M-BART50 [108] and M2M-100 [36] of 418M and 1.2B parameters. The PG models include ruGPT2-Large[20] (760M), ruT5[21] (244M), and mT5 [134] of Small (300M), Base (580M) and Large (1.2B) versions. Each machine-generated sentence undergoes a two-stage annotation procedure on Toloka, as described below.

Data annotation. Each stage includes an unpaid training phase comprising illustrative examples, control tasks for monitoring annotation performance, and the primary annotation phase. The worker is provided with comprehensive instructions that outline the task, clarify the labels, and offer numerous examples. The annotation instructions, examples of the web interface, and other details can be found in §B.1 and §B.2 [70].

[19] github.com/UKPLab/EasyNMT

[20] hf.co/ai-forever/ruGPT2-large

[21] hf.co/cointegrated/rut5-base-paraphraser

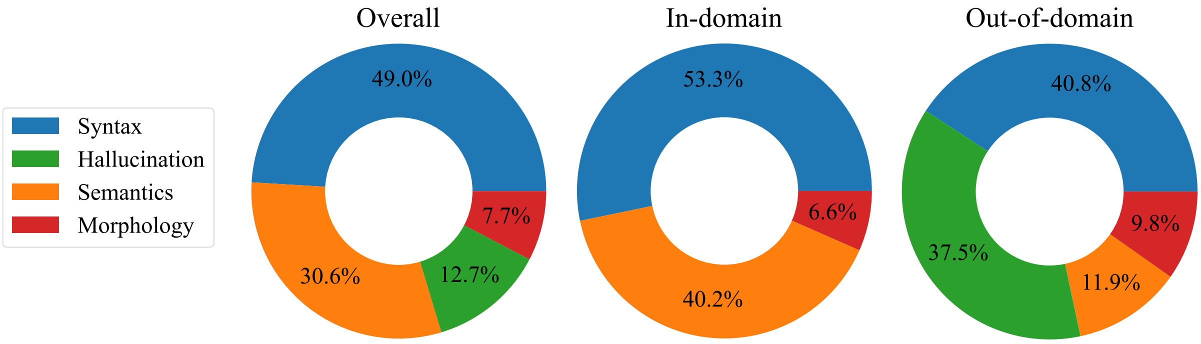


Figure 1: Distribution of violation categories in RuCoLA's unacceptable sentences.

Stage 1: Acceptability judgments. The first annotation stage defines the acceptability of a given sentence. Access to the project is granted to workers certified as native speakers of Russian by Toloka and ranked top 60% workers according to the Toloka rating system. Each worker must answer at least 21 out of 30 training examples correctly. We collect the majority vote labels via a dynamic overlap from three to five workers among those whose annotation quality rate on the control sentences is more than 50%.

Stage 2: Violation categories. The second stage includes validation and annotation of sentences labelled unacceptable in the first stage according to the following categories: "Morphology", "Syntax", "Semantics", "Hallucinations" and "Other". The team of annotators is 30 undergraduate BA and MA students in philology and linguistics from several Russian universities. We hold an online seminar to discuss the annotation task and related works [126; 43; 142]. The students undergo training on 15 examples and can communicate with the paper's authors through a group chat. We keep submissions with more than 30 seconds of response time per five sentences and collect the majority vote labels for each answer option independently. Sentences that fall into multiple violation categories or are labelled "Other" are discarded.

Linguistic phenomena. Each unacceptable sentence falls under one of the four violation categories: morphology, syntax, semantics, and hallucinations. The phenomena are well represented in Russian theoretical and corpus linguistics and are specific to generative models. Examples of the phenomena include incorrect word derivation, agreement violations, corruption of word order, violations of verb transitivity, incorrect use of negation, nonsensical sentences, irrelevant repetitions, decoding confusions, and others (see § A.1 and § A.2 in [70]).

General statistics. The sentences in RuCoLA are filtered by the 4–30 token range with razdel[22]. There are 11 tokens in each sentence on average. Similar to §3.2.1, we estimate the number of high-frequency tokens in each sentence based on the RNC. The sentences contain, on average 92%, of high-frequency tokens. Figure 1 illustrates the distribution of the categories in RuCoLA.

[22] `github.com/natasha/razdel`

Syntactic violations comprise 53.3% and 40.8% in the in-domain and out-of-domain sets. The in-domain set includes 40.2% of semantic and 6.6% of morphological violations, while the out-of-domain set accounts for 11.9% and 9.8%, respectively. 12.7% of the total number of unacceptable sentences are attributed to model hallucinations.

The in-domain set is split into train, development, and private test splits in the 80/10/10 ratio (7.9k/1k/1k examples). The out-of-domain set is divided into development and private test splits in a 50/50 ratio (1.8k/1.8k examples).

3.3.2 Empirical evaluation

Baselines. We evaluate three groups of baselines: non-neural models (TF-IDF and a majority vote classifier), probabilistic acceptability measures (PenLP), and a broad range of monolingual and multilingual Transformer LLMs (ruBERT-base[23], ruRoBERTa-large[24], and ruT5-base[25], XLM-R-base [30], and RemBERT [23]). TF-IDF is a Logistic Regression classifier over TF-IDF features computed on word N-grams with the N-gram range $\in [1;3]$, which results in a total of 2.5k features. PenLP [58] is computed using ruGPT3-medium[26] as the sum of token log-probabilities $P(s)$ normalised by the sentence length with a scaling factor α (see Equation 1). The PenLP approach assigns the label to a given sentence based on the threshold defined on the development set.

$$\text{PenLP}(s) = \frac{P(s)}{((5+|s|)(5+1))^{\alpha}} \tag{1}$$

We conduct a human evaluation on the entire in-domain test set and 50% of the out-of-domain test set. The annotation design is similar to Stage 1: Acceptability judgments, with the exception that (i) we remove the "Not confident" answer option, (ii) the annotators are 16 undergraduate BA and MA students, and (iii) the votes are aggregated using the Dawid-Skene method [32], which is available directly from the Toloka interface. The average quality rate on the control tasks exceeds 75%.

Metrics. The acceptability classification performance is measured by the accuracy score (Acc.) and Matthews Correlation Coefficient (MCC). We train and finetune the baselines by maximising the MCC score on the development set and report the results on the private test set averaged over ten experiment runs from different random seeds.

[23] hf.co/ai-forever/ruBERT-base

[24] hf.co/ai-forever/ruRoBERTa-large

[25] hf.co/ai-forever/ruT5-base

[26] hf.co/ai-forever/ruGPT3-medium

Baseline	Overall		In-domain		Out-of-domain	
	Acc.	MCC	Acc.	MCC	Acc.	MCC
			Non-neural models			
Majority	68.05 ± 0.0	0.0 ± 0.0	74.42 ± 0.0	0.0 ± 0.0	64.58 ± 0.0	0.0 ± 0.0
Linear	67.34 ± 0.0	0.04 ± 0.0	75.53 ± 0.0	0.17 ± 0.0	62.86 ± 0.0	-0.02 ± 0.0
			Acceptability measures from LLMs			
ruGPT-3	55.79 ± 0.0	0.27 ± 0.0	59.39 ± 0.0	0.19 ± 0.0	53.82 ± 0.0	0.30 ± 0.0
			Russian LLMs			
ruBERT	75.9 ± 0.42	0.42 ± 0.01	78.82 ± 0.57	0.4 ± 0.01	74.3 ± 0.71	0.42 ± 0.01
ruRoBERTa	80.8 ± 0.47	0.54 ± 0.01	83.48 ± 0.45	0.53 ± 0.01	79.34 ± 0.57	0.53 ± 0.01
ruT5	71.26 ± 1.31	0.27 ± 0.03	76.49 ± 1.54	0.33 ± 0.03	68.41 ± 1.55	0.25 ± 0.04
			Cross-lingual LLMs			
XLM-R	65.73 ± 2.33	0.17 ± 0.04	74.17 ± 1.75	0.22 ± 0.03	61.13 ± 2.9	0.13 ± 0.05
RemBERT	76.21 ± 0.33	0.44 ± 0.01	78.32 ± 0.75	0.4 ± 0.02	75.06 ± 0.55	0.44 ± 0.01
Human	**84.08**	**0.63**	**83.55**	**0.57**	**84.59**	**0.67**

Table 8: Results for acceptability classification. The best score is in bold, the second best one is underlined.

Acceptability classification

Key results. Table 8 presents the results for acceptability classification. The key findings are: (i) ruRoBERTa and RemBERT achieve the best and second best overall performance among LLMs, which is up to 19 points behind humans in terms of overall MCC score, (ii) the non-neural models receive near-zero performance, (iii) the best-performing LLMs generalise well to the out-of-domain set, with an absolute difference of 0 to 0.04 in terms of MCC, and (iv) the human performance is higher on the out-of-domain set, indicating that unacceptable generated sentences are easier to identify.

Error analysis. The results of manual quantitative analysis of 250 misclassified sentences reveal that (i) classifiers judge ungrammatical sentences with morphological and syntactic violations as acceptable, (ii) humans make mistakes in long sentences with comparative and subordinate clauses and prepositional government, and (iii) most acceptability classifiers achieve high recall on hallucinated sentences, highlighting an application potential for RuCoLA.

Effect of length. The key finding is that the performance is consistent across all methods. However, while the model performance is unstable and slightly degrades as the length increases, the human annotators outperform the LLMs, which is consistent with [124].

Cross-lingual knowledge transfer

RuCoLA contributes to analysing how well the grammatical knowledge in multilingual LLMs is transferred between different languages. We experiment with zero-shot cross-lingual acceptability classification, where the train and development sets are provided in one language and the test data in another one. Here, we use analogous benchmarks in English (CoLA) and Italian (ItaCoLA; [111]) and four multilingual models: mBERT, XLM-R-base, XML-R-large, and RemBERT.

Key results. We observe that (i) there is little difference in performance depending on the source language, (ii) the monolingual scenarios outperform cross-lingual transfer by a large margin, which aligns with the results of [111], (iii) RemBERT performs best in both scenarios among the multilingual LLMs, (iv) RemBERT and XLM-R-large generalise well to the RuCoLA's out-of-domain set when finetuned on English and Italian.

3.3.3 Retrospective

The results on the public leaderboard[27] demonstrate that RuCoLA remains a challenging benchmark for Russian LLMs, with a performance gap of five and eight points of MCC on the in-domain and out-of-domain sets, respectively. More than 30 different approaches to performing acceptability judgments in Russian have been evaluated, including distilled LLMs[28] and TDA-based classifiers [21; 84]. RuCoLA has spurred the practical use of acceptability judgments, e.g., filtering out unacceptable translations of image captions in the ruDALL-E[29] pretraining corpus.

3.4 Findings of the RuATD Shared Task 2022 on Artificial Text Detection in Russian

The LLMs are capable of generating human-like texts among many languages and text domains. However, the LLMs can be misused for malicious purposes [128], e.g., generating fake news [138] and extremist content [68]. The niche field of artificial text detection (ATD) aims to develop resources and computational methods to mitigate the risks of misusing TGMs.

To address this line of research for Russian, we propose the RuATD benchmark, which was organised as a shared task in the framework of the annual "Dialogue" evaluation initiative[30]. RuATD focuses on two task formulations modelled after the Turing test [113] and authorship attribution [114], covering diverse text domains, generative models, and language generation tasks. We host two public leaderboards on the Kaggle competition platform, which remain open to submissions from the community.

[27] rucola-benchmark.com/leaderboard

[28] hf.co/cointegrated/ruBERT-tiny

[29] hf.co/ai-forever/ruDALL-E-Malevich

[30] www.dialog-21.ru/en/evaluation

3.4.1 Method

Detection of neural texts is the first task aimed at predicting whether a given text is generated or written by a human. The task is framed as a binary classification problem.

- **Text:** *"Я был готов помочь ему в опасности своей жизни."*
- **Answer:** Machine

Authorship attribution is the second task aimed at identifying the author of a given text. The task is a multi-class classification problem with 14 target classes: a human and 13 models.

- **Text:** *"Я был полон решимости помочь ему, даже рискуя собственной жизнью."*
- **Answer:** Human

Data collection. RuATD is composed of 215k human-written and artificial texts generated with 13 models. The methodology for creating RuATD includes three main steps: (i) collecting human-written texts, (ii) text generation, and (iii) post-processing and filtering. We provide a brief description below and refer the reader to §2 in [103] for details on model and data specification, model finetuning hyperparameters, and text generation hyperparameters.

1. **Human-written text curation:** We collect human-written texts from task-specific datasets and publicly available resources among multiple text domains: RNC, social media posts, Wikipedia articles (top-100 most viewed pages in 2016-2021), news articles (Lenta.ru, KP, Interfax, Izvestia, and others), digitalised diaries [69], and strategic documents from the Ministry of Economic Development of the Russian Federation [48]. We also collect gold standard translation references from WikiMatrix and Tatoeba.
2. **Artificial text generation:** We use human-written texts as the input to the generative models finetuned for one or more language generation tasks: MT, PG, text simplification, and text summarisation. In addition, we consider back-translation and open-ended generation.

   - **MT & back-translation:** We translate subsets of Tatoeba and WikiMatrix in three language pairs (English/French/Spanish-Russian) with three models from the EasyNMT library: OPUS-MT, M-BART50, and M2M-100. In the back-translation setup, the sentence in Russian is translated into a target language and then translated back into Russian.
   - **PG:** paraphrases are generated by models via the russian-paraphrasers library: ruGPT2-large, ruT5-base, and mT5 of Small and Large versions.
   - **Text simplification:** We finetune ruGPT3-small, ruGPT3-medium, ruGPT3-large, mT5-large, and ruT5-large on the RuSimpleSentEval dataset [95].
   - **Text summarisation:** We use two summarisation models finetuned on Gazeta [41]: ruT5-base and M-BART.

Task	Model	Size	N	%	Domain	Task	Model	Size	N	%	Domain
Back-translation	Human M-BART50 M2M-100 OPUS-MT	35,588	12.9	88.0	RNC, Wikipedia, news, diaries, WikiMatrix, Tatoeba, SD	**Machine translation**	Human M-BART50 M2M-100 OPUS-MT	35,860	11.5	89.0	WikiMatrix, Tatoeba
Open-ended generation	Human ruGPT3-small ruGPT3-medium ruGPT3-large	37,499	141.5	85.0	RNC, Wikipedia, news, diaries, SD, social media	**Text summarisation**	Human M-BART M-BART50 ruT5-base	17,164	33.5	86.0	RNC, Wikipedia, news, diaries, SD
Paraphrase generation	Human mT5-small mT5-large ruGPT2-large ruGPT3-large ruT5-base	44,298	13.0	85.0	RNC, SD, social media, Wikipedia, news, diaries	**Text simplification**	Human mT5-large ruGPT3-small ruGPT3-medium ruGPT3-large ruT5-large	44,700	18.3	86.0	RNC, SD, social media, Wikipedia, news, diaries

Table 9: General statistics of the RuATD benchmark. **Notations**: N=average number of tokens; %=percentage of high-frequency tokens; SD=strategic documents; RNC=Russian National Corpus.

- **Open-ended generation:** We generate texts in a zero-shot manner by prompting the ruGPT3-small, ruGPT3-medium, and ruGPT3-large models with the text beginning.

3. **Post-processing & filtering:** We use a set of heuristics to (i) discard text duplicates, copied inputs, empty outputs, (ii) discard texts containing obscene lexis, (iii) keep translations classified as Russian by the language detection model[31], and (iv) filter the texts by the token range: 5-to-25 (MT, back-MT, PG), 10-to-30 (text simplification), 15-to-60 (text summarisation), and 85-to-400 (open-ended generation).

General statistics. Table 9 presents the general statistics of the RuATD benchmark by language generation task, generative model, and text domain. There are 37.9 tokens on average, with variations depending on the language generation task. The fraction of high-frequency tokens based on the RNC is similar among the human-written and machine-generated texts: 86% and 87%, respectively. We split the corpus into four sets in the 60/10/15/15 proportion ratio: train (130k), development (21k), public test (32k), and private test (32k). The public test set is available during the competition, allowing the participants to develop and improve their submissions. The private test set defines the final rankings of the participants, preventing overfitting on the public test set. The train, development, and public and private test sets are utilised in both task formulations, with the only difference in the target classes. Specifically, the machine label in the binary classification task is broken into 13 model names in the multi-class classification task.

3.4.2 Empirical evaluation

Baselines. We provide two baseline solutions to the shared task participants: TF-IDF and ruBERT-base. TF-IDF refers to a Logistic Regression classifier over TF-IDF features computed on word N-grams with the N-gram range $\in [1;3]$. The feature dimensionality is reduced with

[31] github.com/fedelopez77/langdetect

Rank	Detection of neural texts		Authorship attribution	
	Team	Acc.	Team	Acc.
1	MSU	0.829	Posokhov Pavel	0.650
2	Igor	0.827	Yixuan Weng	0.647
3	Orzhan	0.826	Orzhan	0.646
4	mariananieva	0.824	MSU	0.628
5	Ivan Zakharov	0.822	ruBERT baseline	0.598
6	Yixuan Weng	0.818	Nikita Selin	0.590
7	ilya koziev	0.817	Victor Krasilnikov	0.550
8	miso soup	0.811	Petr Grigoriev	0.458
9	Eduard Belov	0.810	TF-IDF baseline	0.443

Table 10: Top-nine positions on the RuATD leaderboards.

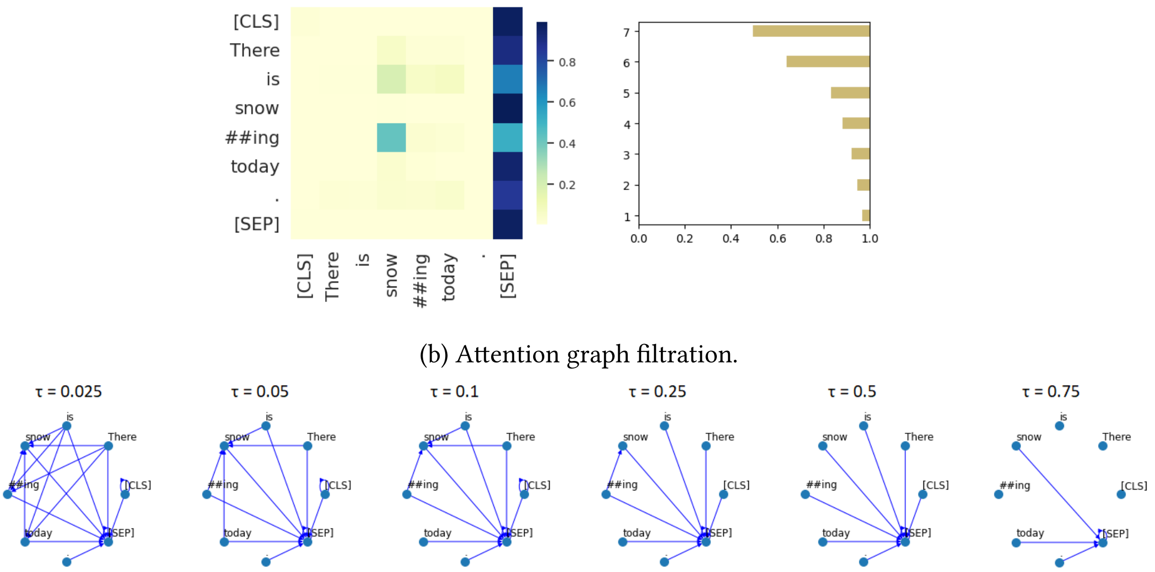
Singular Value Decomposition to 5k. ruBERT-base is fine-tuned on the corresponding task. We also establish a human baseline for detection of neural texts using stratified subsets of 2.5k samples from the public and private tests. § A in [103] presents the annotation instruction given to crowdsourcing workers on Toloka. We grant access to the project to workers ranked top 70% . Each worker must finish the training stage by answering at least 27 out of 32 examples correctly. We aggregate the majority vote labels via dynamic overlap from three to five trained workers after (i) discarding votes from workers whose annotation quality rate on the control tasks is less than 50% and (ii) filtering out submissions with less than 15 seconds of response time per five texts.

Metrics. We use the accuracy score to evaluate systems in each task.

Key results. We outline the key findings of evaluating the baselines and 38 shared task solutions (see Table 10): (i) the performance of the detectors depends on the text length (the longer the text, the higher the accuracy), (ii) identifying the author of the given text is not trivial, meaning that human-written texts and texts generated by different models have similar properties, and (iii) humans underperform the systems on detection of neural texts by up to 0.169 accuracy score, which is consistent with [51; 115].

3.4.3 Retrospective

The ATD field remains focused on three languages: English, Chinese, and Russian. The ATD benchmarks are becoming more complex, covering various domains, architectures of generative LLMs, decoding methods, and strategies for generating texts. The rapid proliferation of generative LLMs necessitates continuous updating of the ATD benchmarks and the development of more generalisable artificial text detectors since the detection performance decreases with the LLMs' scaling [106]. In §3.5, we present a novel neural text detector, which outperforms existing detectors, and demonstrates more robust generalisation to unseen larger GPT-2 models.



(a) Attention map (left); Barcode (right).

(b) Attention graph filtration.

Figure 2: An example of attention maps, barcodes, and filtration procedure [21].

3.5 Artificial Text Detection via Examining the Topology of Attention Maps

With recent advances in language generation, the problem of low human performance in detecting neural texts becomes more prominent, stimulating the development of reliable neural text detectors to assist humans. Different computational approaches have been proposed, such as training linear classifiers over count-based and linguistic features [47; 9], utilising statistical properties computed by pretrained Transformer LLMs [40; 34], and finetuning Transformer LLMs [106]. The Transformer-based detectors are highly effective in ATD tasks but lack interpretability and robustness towards unseen generative models [49].

We introduce a hybrid artificial text detector that combines the advantages of feature-based and Transformer-based detectors – interpretability and strong performance. We make one of the first attempts to adapt methods from applied topology and computational geometry to the Transformer's attention mechanism. Our approach includes (i) extracting three types of TDA features from a graph representation of the Transformer's attention map and (ii) training a linear classifier over the concatenation of the features.

3.5.1 Method

We treat the Transformer's attention map (see Figure 2a; left) as a weighted graph, where the vertices are text tokens, and the edges' weights correspond to the attention weights. This representation is used to obtain *"filtration"*, i.e., an ordered set of attention graphs filtered by increasing attention weight thresholds τ_i (see Figure 2b). Filtering edges lower than the given threshold affects the graph structure and features. TDA techniques track these changes, identifying

Domain	Model	|Train|		|Dev|		|Test|		|Vocab|		Length	
		H	M	H	M	H	M	H	M	H	M
Reddit	GPT-2-small pure sampling	20k	20k	2.5k	2.5k	2.5k	2.5k	220k	532k	593 ± 177	515 ± 322
Amazon reviews	GPT-2-XL pure sampling	5k	5k	1k	1k	4k	4k	47k	49k	179 ± 170	177 ± 171
RealNews	GROVER top-p sampling	5k	5k	1k	1k	4k	4k	98k	75k	721 ± 636	519 ± 203

Table 11: Statistics for the datasets used in the experiments on the artificial text detection task. **Notations**: **H**=Human; **M**=Machine.

the moments when the features appear (i.e., their "*birth*") and disappear (i.e., their *"death"*). The feature's lifetime is encoded as a set of intervals called a *"barcode"* (see Figure 2a; right), where each interval (*"bar"*) lasts from the feature's birth to its death. The barcode characterises the stability of the graph features.

We extract three types of TDA features from the Transformer's attention maps. The features are computed over pre-defined thresholds using each attention head and further concatenated.

1. **Topological features**. Topological features include the first two Betti numbers of the undirected graph β_0 and β_1 and standard properties of the directed graph, such as the number of strongly connected components, edges, and cycles.
2. **Features derived from barcodes.** Barcode is the representation of the graph's persistent homology. We compute 0/1-dimensional barcodes from the attention graph and their descriptive characteristics: the sum/average/variance of lengths of bars, the number of bars with the time of birth/death greater/lower than a threshold, and the entropy of the barcodes.
3. **Features based on the distance to patterns.** The shape of attention graphs has various attention patterns: attention to the previous/current/next token, attention to the [SEP]/[CLS] token, and attention to punctuation marks [25]. Figure 2b shows an example of the attention to the [SEP] token pattern. We represent attention patterns as binary matrices and calculate the Frobenius norm of the difference between the matrices normalised by the sum of their norms.

3.5.2 Empirical evaluation

Datasets. We conduct the experiments using subsets of three datasets from different domains (see Table 11): (i) Reddit & GPT-2-small [86], (ii) Amazon reviews & GPT-2-XL [4; 106], and (iii) RealNews & GROVER [138].

Model	Reddit & GPT-2 Small	Amazon Reviews & GPT-2 XL	RealNews & GROVER
TF-IDF, N-grams	68.1	54.2	56.9
BERT [CLS trained]	77.4	54.4	53.8
BERT [Fully finetuned]	**88.7**	**60.1**	**62.9**
BERT [SLOR]	78.8	59.3	53.0
Topological features	86.9	59.6	63.0
Features derived from barcodes	84.2	60.3	61.5
Features based on the distance to patterns	85.4	61.0	62.3
All features	**87.7**	**61.1**	**63.6**

Table 12: The results of the artificial text detection experiments. The values are scaled by 100.

Baselines. The BERT-based[32] baselines include: (i) BERT [CLS trained] is a linear layer trained over [CLS]-pooled text representations, with the BERT's weights frozen, (ii) BERT [Fully finetuned] is a finetuned BERT model. In addition, we train a Logistic Regression classifier over (iii) TF-IDF N-grams with the N-gram range $\in [1, 2]$ and (iv) BERT [SLOR] [82], a pseudo-perplexity-based acceptability measure [58].

Models. We train a Logistic Regression classifier over TDA features derived from the attention matrices from the BERT model: (i) *Topological features*, (ii) *Features derived from barcodes*, (iii) *Features based on distance to patterns*, and (iv) *All features*, which is the concatenation of all features. The performance is evaluated with the accuracy score.

Detection of neural texts

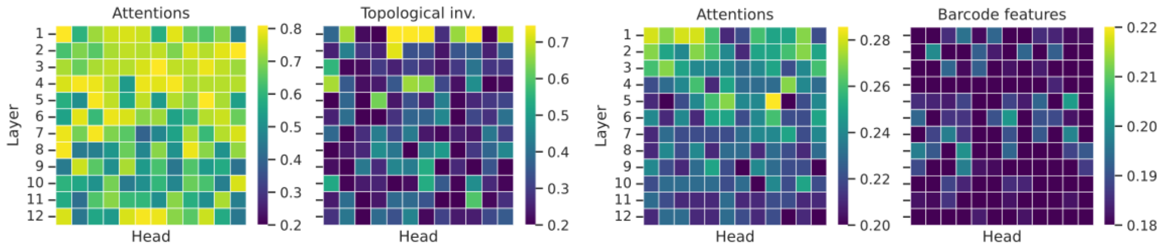
Key results. Table 12 presents the results for detecting artificial texts. The key findings are that the TDA-based detectors (i) outperform the count-based and BERT-based baselines by up to 10% accuracy score, and (ii) achieve performance comparable with the finetuned BERT.

Robustness to unseen generative models. Here, the detectors are trained on human-written texts and texts generated by the smallest GPT-2 and used to detect texts generated by unseen GPT-style models with pure sampling: GPT-2-medium (345M), GPT-2-large (762M) and GPT-2-xl (1542M). The key finding is that detector trained *Topological features* is more generalisable than the baselines but performs slightly worse than the finetuned BERT on the GPT-2-small test set.

Probing analysis

Key results. Figure 3 presents the results of the probing analysis on two probing tasks: predicting the sentence length and the depth of the syntax tree. We find that the TDA features (i) are sensitive to the properties, (ii) can lose information about properties, which is indicated

[32] hf.co/bert-base-uncased



(a) Sentence length

(b) Depth of the syntax tree

Figure 3: Heat maps of attention head-wise probin. Attentions=Frozen attention weights. X-axis=Head index number. Y-axis=Layer index number. The brighter the color, the higher the accuracy score for the attention head.

by different performance scores for the same attention heads, and (iii) do not encode semantic properties (e.g., verb tense), but this information is sufficient for the ATD task.

3.5.3 Retrospective

TDA has found broad application in different ML tasks, such as human action recognition [107], image segmentation [26], text classification [98; 129] and language generation evaluation [28]. Multiple follow-up works have reported that our methodology can be adapted to promote state-of-the-art results on speech processing tasks [112] and reach the human-level performance on linguistic acceptability tasks [21].

3.6 Vote'n'Rank: Revision of Benchmarking with Social Choice Theory

The appropriateness of the arithmetic mean aggregation procedure in multi-task ML benchmarks is questioned for its properties: (i) implying that all task metrics are homogeneous [29], (ii) ignoring task complexity [71], (iii) relying on the absolute score difference [83], and (iv) ranking systems higher when they outperform the others only on the outlier tasks [1].

To address these limitations, we introduce Vote’n’Rank, an alternative tool for ranking NLP systems in multi-task and multi-criteria evaluation setups based on the social choice theory [2]. Vote’n’Rank includes eight interpretable aggregation procedures that rely on rankings in each criterion (e.g., task performance, computational efficiency, and fairness) and allow aggregating heterogeneous information. This section briefly describes the aggregation procedures and experimental design and outlines the key findings.

3.6.1 Method

We adopt the conceptual definitions from the social choice theory to the objectives of selecting the best-performing system and ranking a set of systems as follows: *(i) a voter* or a *criterion* is a task in a given benchmark, and *(ii) an alternative* is a system candidate.

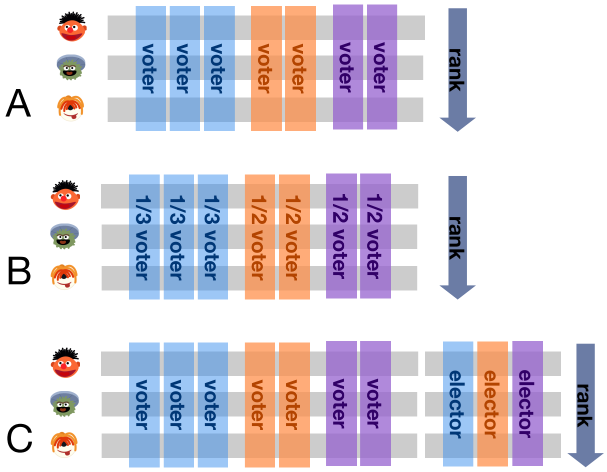


Figure 4: Three ways to run the aggregation procedures. A: Basic aggregation. B: Weighted aggregation. C: Two-step aggregation.

Aggregation procedures. The Vote'n'Rank aggregation procedures belong to three classes of voting rules: scoring rules (*Plurality*, *Borda*, and *Dowdall*), iterative scoring rules (*Threshold* and *Baldwin*), and majority-relation based rules (*Condorcet*, *Copeland*, and *Minimax*). The rules' design is based on the mathematical foundations of the social choice theory and is generally accepted in the community [3; 6; 76]. We offer recommendations for choosing the rule below and refer the reader to §2 and § A for a formal description of the rules and their properties, and illustrative examples on how they work.

- The *Plurality* rule is a good choice if the user wants only the best systems in each criterion.
- *Borda* and *Dowdall* are most suitable when all ranking positions matter. Note that *Dowdall* assigns higher weights to the top positions.
- The *Threshold* rule is helpful in cases when the user wants to minimise the number of the low-performance criteria: the rule assigns the highest rank to the alternative that is considered the worst in the least amount of criteria.
- The *Baldwin*, *Condorcet*, *Copeland*, or *Minimax* rules should be used when the goal is to select the alternative that beats all the others in pairwise comparison. The main difference is how the rules behave when there is no such alternative. In particular, *Baldwin* selects the alternative that is left after elimination according to the *Borda* scores. *Copeland* chooses the alternative that dominates the others in more cases and is dominated by the least. In turn, *Minimax* selects the alternative with minimum defeat in pairwise comparison.

Framework. Figure 4 describes three scenarios to run the aggregation procedures. The toy benchmark has three evaluated alternatives and consists of seven voters grouped by the task, e.g., MRC, text classification, and question answering.

A Basic aggregation: the aggregation procedure is applied to the leaderboard as is.

B Weighted aggregation: each voter in the group is assigned a group weight equal to $1/|T_{group}|$. The blue group weights are $1/3$, and the orange and the violet group weights are $1/2$. Each group contributes equally to the final ranking, regardless of the number of voters.

C Two-step aggregation: each voter group is treated as a standalone leaderboard. We independently apply a procedure to each voter group and compute an interim ranking shown as "elector". Next, we aggregate the group-wise rankings by applying the same procedure one more time and compute the final ranking.

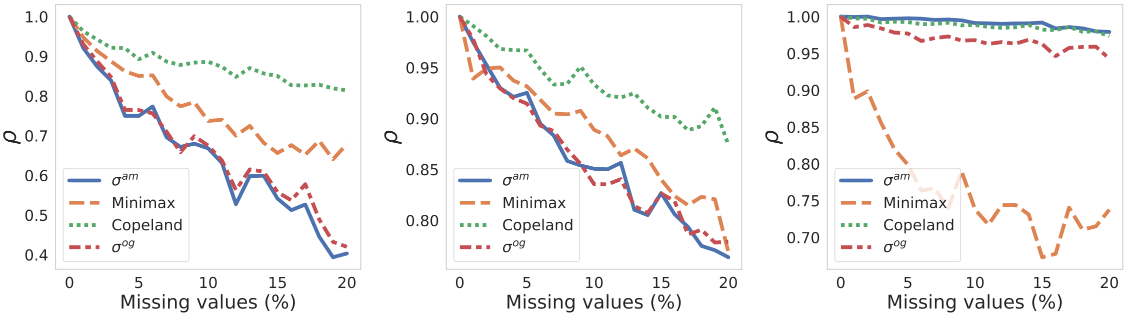
3.6.2 Empirical evaluation

We design four application-oriented case studies to compare our framework with three rank aggregation procedures: (i) σ^{am} is the arithmetic mean aggregation procedure, (ii) σ^{gm} is the geometric mean aggregation procedure, and (iii) σ^{og} [1] is an aggregation metric that identifies the amount by which the system fails to get a minimum score of $\gamma = 0.95$ (lower is better). The experiments are conducted using the GLUE, SuperGLUE, and VALUE public leaderboards.

Re-interpreting benchmarks

Case study description. The first case study considers two experiment settings: (i) re-ranking systems on the leaderboards using the scoring and majority-relation based rules, and (ii) selecting the single-winner systems using all rules. We compare the rankings with the baselines by computing *(i)* the agreement rate (AR; in %), i.e., the intersection of the top/least-k systems between the given procedure and σ^{am}, *(ii)* the Kendall Tau correlation (τ) between the total rankings, *(iii)* the discriminative power (DP) or the number of equivalent alternatives [16], and *(iv)* the independence of irrelevant alternatives (IIA), i.e., how often the new system changes the ranking.

Key results. We find that (i) the procedures agree with one another on particular top/least-k systems, but output different rankings, (ii) *Dowdall* and *Borda* produce zero ore only one tied alternative, while *Plurality* and *Minimax* treat a larger number of systems equivalent, (iii) *Copeland*, *Minimax*, and *Plurality* are least impacted by introducing a new system, (iv) HUMAN may still take leading positions on GLUE according to *Plurality* and *Dowdall*, since HUMAN annotators receive the best performance on RTE [120] and MNLI [131], (v) HUMAN is declared the winner on SuperGLUE by the *Copeland*, *Plurality*, and *Dowdall* procedures, and (vi) HUMAN is ranked first on VALUE by the *Copeland*, *Minimax*, and *Condorcet* procedures, meaning that HUMAN beats any other model in pairwise comparison.



(a) GLUE

(b) SuperGLUE

(c) VALUE

Figure 5: Spearman correlation (ρ) between top-7 model rankings with and without omitted leaderboard values.

The Condorcet winner

Case study description. The Condorcet method declares a system the winner if it dominates all other alternatives in pairwise comparison [14]. In practice, the method is hard to destabilise and easy to interpret. In the second case study, we aim to find the weight vector which makes a given alternative the Condorcet winner or determine that no such weights exist.

Key results. There are nine, ten, and three alternatives that can be declared the Condorcet winners in specific evaluation protocols on GLUE, SuperGLUE, and VALUE, respectively. The existence of the Condorcet winner weight vectors allows practitioners to simulate real-world scenarios where the system is the best across the given evaluation criteria.

Robustness to omitting scores

Case study description. The third case study provides a more detailed analysis of the majority-relation based procedures that handle missing performance scores. We evaluate robustness of *Copeland*, *Minimax*, σ^{am}, and σ^{og} to omitting scores. First, we compute the rankings on each benchmark without omitting scores and use them as references. Next, we randomly replace N scores with empty values and find top-7 systems over the corrupted leaderboards. We calculate the Spearman correlation (ρ) between the final rankings and the references. Note that we use the median values when omitting scores for σ^{am} and σ^{og} as the baselines. The results are averaged over 100 experiment runs.

Key results. Figure 5 presents the results of the experiment. σ^{am} and σ^{og} demonstrate lower stability, and *Copeland* performs the best on GLUE and SGLUE. However, we observe that *Minimax* is the least stable on VALUE, whilst *Copeland*, σ^{am}, and σ^{og} perform on par. We attribute the low stability of *Minimax* on VALUE to its limitations. There are minor differences between the alternatives on VALUE, which cause *Minimax* to score them very similar and be sensitive to any missing value.

Ranking based on user preferences

Case study description. The fourth case study aims at ranking NLP systems based on the user utility. We simulate a practical scenario, where the user preferences include task performance, computational efficiency, and fairness. In this experiment, we consider BERT-base, RoBERTa-base, ALBERT-base, DeBERTa-base [44], DistilBERT-base [97], DistilRoBERTa-base [97], and GPT2-medium. We finetune and evaluate the LLMs on the GLUE benchmark, estimate their computational efficiency during finetuning via the Impact tracker toolkit [45], and measure fairness on three social bias evaluation datasets: CrowS-Pairs [74], StereoSet [73], and Winobias [141].

Key results. We report the key findings when using the *Borda* procedure, with the weights vector $(0.4, 0.3, 0.3)$ assigned to performance, computational efficiency, and fairness. The distilled LLMs (DistilRoBERTa and DistilBERT) are declared the winners, while the best-performing LLMs (e.g., DeBERTa) is ranked in the middle positions due to sub-optimal computational efficiency and satisfying performance on detecting social biases. Comparing our results with the Dynascore [66], we find that the average performance ranking is not preserved when using our voting rules. This is due to the fact that Dynascore assigns a weight of 0.5 to performance, which blocks substantial changes in re-ranking.

3.6.3 Retrospective

The nuanced question of how to aggregate results in multi-task benchmarks applies to each proposed standardised evaluation resource. Vote'n'Rank allows re-interpreting saturated benchmarks, which undergo stagnation in improvements of the state-of-the-art results after surpassing or reaching the human-level performance. The problem of benchmark saturation is widely discussed in the NLP community, particularly in light of the state-of-the-art chasing tendencies with minor performance gains [91; 88; 54]. While the criticism is of utmost importance, it relies on the conclusions drawn from utilising the arithmetic mean aggregation procedure. Vote'n'Rank highlights that – at the moment of writing the publication – humans still outperform the LLMs when accounting for preferences in each task.

The comparison of systems with Vote'n'Rank is hindered by the absence of the correct ranking, specifically when performances are incomplete (e.g., the absence of human performance scores on the VALUE text-to-video retrieval and video captioning tasks[33]). However, we make application-oriented contributions, offering alternative aggregation procedures for evaluating systems irrespective of the ML area. Our framework is relevant in view of the emerging benchmarking paradigm, which aims evaluate LLMs exhaustively on user-oriented scenarios [62].

[33] value-benchmark.github.io/leaderboard.html

4 Conclusion

The final section summarises contributions of this thesis. The benchmarks, source code, leaderboards, human evaluation projects, and other materials are publicly available under the Apache 2.0 license.

1. We propose the Russian SuperGLUE, RuCoLA, and RuATD benchmarks, which have become standardised evaluation resources for measuring the advancement of Russian LLMs. The benchmarks cover 11 diverse NLU tasks in various formulations, including machine reading comprehension, question answering, word sense disambiguation, natural language inference, coreference resolution, acceptability classification, detection of neural texts, and authorship attribution. We establish methodologies for evaluating human annotators and collecting and annotating data, which account for specifics of the Russian language. Each benchmark provides a public leaderboard for comparing the results of the state-of-the-art LLMs against the human level.
2. We introduce a novel TDA-based artificial text detector that utilises three types of interpretable features extracted from the graph representation of the Transformer's attention maps. The features display sensitivity to the surface and syntactic properties of the text. Our approach outperforms existing detectors on three datasets from different text domains and demonstrates more robust generalisation to unseen GPT-2 LLMs.
3. We develop Vote'n'Rank, a flexible framework for ranking and determining single-winner LLMs with eight interpretable aggregation procedures stemming from the social choice theory. Vote'n'Rank addresses the fundamental limitations of the arithmetic mean aggregation procedure in multi-task benchmarks and multi-criteria evaluation setups. We offer recommendations based on the procedures' properties and the intended application scenarios of the framework.
4. We conduct a detailed empirical evaluation of more than 100 LLMs and their configurations using the proposed evaluation resources and tools. The experiment results show that (i) the LLMs fall behind humans by a large margin on most of the NLU tasks, while humans significantly underperform the LLMs on the detection of neural texts, (ii) the LLMs for Russian struggle to judge unacceptable sentences with morphological and semantic violations but generalise well to machine-generated sentences, (iii) the cross-lingual knowledge transfer on acceptability classification across Russian, English, and Italian is challenging, (iv) the best-performing LLMs on standardised benchmarks are less preferred when accounting for their computational efficiency and fairness, and (v) humans can take the leading positions on the saturated benchmarks according to the Vote'n'Rank procedures since they still receive the best performance on particular tasks.

References

[1] Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. Deep Reinforcement Learning at the Edge of the Statistical Precipice. *Advances in Neural Information Processing Systems*, 34, 2021.

[2] Mark Aizerman and Fuad Aleskerov. *Theory of Choice*, volume 38. North Holland, 1995.

[3] Fuad Aleskerov, Vyacheslav V Chistyakov, and Valery Kalyagin. The threshold aggregation. *Economics Letters*, 107(2):261–262, 2010.

[4] Amazon. Amazon Customer Reviews Dataset. `https://s3.amazonaws.com/amazon-reviews-pds/readme.html`, 2019.

[5] Alexandra Antonova and Alexey Misyurev. Building a web-based parallel corpus and filtering out machine-translated text. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 136–144, Portland, Oregon, June 2011. Association for Computational Linguistics.

[6] Kenneth J Arrow. Social Choice and Individual Values. In *Social Choice and Individual Values*. Yale university press, 2012.

[7] Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online, July 2020. Association for Computational Linguistics.

[8] Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 2019.

[9] Sameer Badaskar, Sachin Agarwal, and Shilpa Arora. Identifying real or fake articles: Towards better language modeling. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*, 2008.

[10] Shreyan Bakshi, Soumya Batra, Peyman Heidari, Ankit Arun, Shashank Jain, and Michael White. Structure-to-text generation with self-training, acceptability classifiers and context-conditioning for the GEM shared task. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 136–147, Online, August 2021. Association for Computational Linguistics.

[11] Soumya Batra, Shashank Jain, Peyman Heidari, Ankit Arun, Catharine Youngs, Xintong Li, Pinar Donmez, Shawn Mei, Shiunzu Kuo, Vikas Bhardwaj, Anuj Kumar, and Michael White. Building adaptive acceptability classifiers for neural NLG. In *Proceedings of the*

2021 Conference on Empirical Methods in Natural Language Processing, pages 682–697, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[12] Anya Belz, Anastasia Shimorina, Shubham Agarwal, and Ehud Reiter. The ReproGen Shared Task on Reproducibility of Human Evaluations in NLG: Overview and Results. In *The 14th International Conference on Natural Language Generation*, 2021.

[13] Emily Bender. The BenderRule: On Naming the Languages We Study and Why It Matters. *The Gradient*, 2019.

[14] Duncan Black et al. The theory of committees and elections. 1958.

[15] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258*, 2021.

[16] Felix Brandt and Hans Georg Seedig. On the Discriminative Power of Tournament Solutions. In *Operations Research Proceedings 2014*, pages 53–58. Springer, 2016.

[17] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language Models are Few-shot Learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[18] Franck Burlot, Yves Scherrer, Vinit Ravishankar, Ondřej Bojar, Stig-Arne Grönroos, Maarit Koponen, Tommi Nieminen, and François Yvon. The WMT'18 morpheval test suites for English-Czech, English-German, English-Finnish and Turkish-English. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 546–560, Belgium, Brussels, October 2018. Association for Computational Linguistics.

[19] Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yuri Kuratov, Denis Kuznetsov, Alexey Litinsky, Varvara Logacheva, Alexey Lymar, Valentin Malykh, Maxim Petrov, Vadim Polulyakh, Leonid Pugachev, Alexey Sorokin, Maria Vikhreva, and Marat Zaynutdinov. DeepPavlov: Open-source library for dialogue systems. In *Proceedings of ACL 2018, System Demonstrations*, pages 122–127, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[20] Frédéric Chazal and Bertrand Michel. An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists. *Frontiers in artificial intelligence*, 4:667963, 2021.

[21] Daniil Cherniavskii, Eduard Tulchinskii, Vladislav Mikhailov, Irina Proskurina, Laida Kushnareva, Ekaterina Artemova, Serguei Barannikov, Irina Piontkovskaya, Dmitri Piontkovski, and Evgeny Burnaev. Acceptability judgements via examining the topology of attention maps. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 88–107, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[22] Noam Chomsky. *Aspects of the Theory of Syntax*. MIT Press, 1965.

[23] Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. Rethinking Embedding Coupling in Pre-trained Language Models. In *International Conference on Learning Representations*, 2020.

[24] Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470, 2020.

[25] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy, August 2019. Association for Computational Linguistics.

[26] James Clough, Nicholas Byrne, Ilkay Oksuz, Veronika A Zimmer, Julia A Schnabel, and Andrew King. A Topological Loss Function for Deep Learning-based Image Segmentation Using Persistent Homology. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[27] Pierre Colombo, Nathan Noiry, Ekhine Irurozki, and Stephan Clemencon. What are the Best Systems? New Perspectives on NLP Benchmarking. In *Advances in Neural Information Processing Systems*, 2022.

[28] Pierre Colombo, Guillaume Staerman, Chloé Clavel, and Pablo Piantanida. Automatic text evaluation through the lens of Wasserstein barycenters. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10450–10466, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[29] Pierre Jean A Colombo, Chloé Clavel, and Pablo Piantanida. InfoLM: A New Metric to Evaluate Summarization & Data2Text Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10554–10562, 2022.

[30] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin

Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics.

[31] Jurafsky Daniel, Martin James H, et al. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 2007.

[32] A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28, 1979.

[33] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[34] Liam Dugan, Daphne Ippolito, Arun Kirubarajan, and Chris Callison-Burch. RoFT: A tool for evaluating human detection of machine-generated text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 189–196, Online, October 2020. Association for Computational Linguistics.

[35] Pavel Efimov, Andrey Chertok, Leonid Boytsov, and Pavel Braslavski. SberQuAD–Russian Reading Comprehension Dataset: Description and Analysis. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings 11*, pages 3–15. Springer, 2020.

[36] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. Beyond English-centric Multilingual Machine Translation. *Journal of Machine Learning Research*, 22(107):1–48, 2021.

[37] Alena Fenogenova. Russian paraphrasers: Paraphrase with transformers. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 11–19, Kiyv, Ukraine, April 2021. Association for Computational Linguistics.

[38] Alena Fenogenova, Vladislav Mikhailov, and Denis Shevelev. Read and reason with MuSeRC and RuCoS: Datasets for machine reading comprehension for Russian. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6481–6497, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[39] Alena Fenogenova, Tatiana Shavrina, Alexandr Kukushkin, Maria Tikhonova, Anton Emelyanov, Valentin Malykh, Vladislav Mikhailov, Denis Shevelev, and Ekaterina Artemova. Russian SuperGLUE 1.1: Revising the Lessons not Learned by Russian NLP-models. 2021.

[40] Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy, July 2019. Association for Computational Linguistics.

[41] Ilya Gusev. Dataset for automatic summarization of russian news. In *Artificial Intelligence and Natural Language*, pages 122–134. Springer International Publishing, 2020.

[42] Mareike Hartmann, Miryam de Lhoneux, Daniel Hershcovich, Yova Kementchedjhieva, Lukas Nielsen, Chen Qiu, and Anders Søgaard. A multilingual benchmark for probing negation-awareness with minimal pairs. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 244–257, Online, November 2021. Association for Computational Linguistics.

[43] Jie He, Bo Peng, Yi Liao, Qun Liu, and Deyi Xiong. TGEA: An error-annotated dataset and benchmark tasks for TextGeneration from pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6012–6025, Online, August 2021. Association for Computational Linguistics.

[44] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*, 2020.

[45] Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning. *Journal of Machine Learning Research*, 21(248):1–43, 2020.

[46] Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. XTREME: A massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR, 2020.

[47] Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822, Online, July 2020. Association for Computational Linguistics.

[48] Vitaly Ivanin, Ekaterina Artemova, Tatiana Batura, Vladimir Ivanov, Veronika Sarkisyan, Elena Tutubalina, and Ivan Smurov. Rurebus-2020 shared task: Russian relation extraction for business. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference Dialogue* [*Kompiuternaia Lingvistika i Intellektualnye Tehnologii: Trudy Mezhdunarodnoj Konferentsii Dialogue*], Moscow, Russia, 2020.

[49] Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks Lakshmanan, V.S. Automatic detection of machine generated text: A critical survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2296–2309, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[50] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online, July 2020. Association for Computational Linguistics.

[51] Marzena Karpinska, Nader Akoury, and Mohit Iyyer. The perils of using Mechanical Turk to evaluate open-ended text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265–1285, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[52] Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[53] Daniel Khashabi, Arman Cohan, Siamak Shakeri, Pedram Hosseini, Pouya Pezeshkpour, Malihe Alikhani, Moin Aminnaseri, Marzieh Bitaab, Faeze Brahman, Sarik Ghazarian, Mozhdeh Gheini, Arman Kabiri, Rabeeh Karimi Mahabagdi, Omid Memarrast, Ahmadreza Mosallanezhad, Erfan Noury, Shahab Raji, Mohammad Sadegh Rasooli, Sepideh Sadeghi, Erfan Sadeqi Azer, Niloofar Safi Samghabadi, Mahsa Shafaei, Saber Sheybani, Ali Tazarv, and Yadollah Yaghoobzadeh. ParsiNLU: A suite of language understanding challenges for Persian. *Transactions of the Association for Computational Linguistics*, 9:1147–1162, 2021.

[54] Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in NLP.

In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online, June 2021. Association for Computational Linguistics.

[55] Samuel Kounev, Klaus-Dieter Lange, and Jóakim von Kistowski. *Systems Benchmarking: For Scientists and Engineers*, volume 1. Springer, 2020.

[56] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese general language understanding evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2957–2966, Marseille, France, June 2022. European Language Resources Association.

[57] Laida Kushnareva, Daniil Cherniavskii, Vladislav Mikhailov, Ekaterina Artemova, Serguei Barannikov, Alexander Bernstein, Irina Piontkovskaya, Dmitri Piontkovski, and Evgeny Burnaev. Artificial text detection via examining the topology of attention maps. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 635–649, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[58] Jey Han Lau, Carlos Armendariz, Shalom Lappin, Matthew Purver, and Chang Shu. How furiously can colorless green ideas sleep? sentence acceptability in context. *Transactions of the Association for Computational Linguistics*, 8:296–310, 2020.

[59] Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online, November 2020. Association for Computational Linguistics.

[60] Byron C Lewis and Albert E Crews. The Evolution of Benchmarking as a Computer Performance Evaluation Technique. *MIS Quarterly*, pages 7–16, 1985.

[61] Linjie Li, Jie Lei, Zhe Gan, Licheng Yu, Yen-Chun Chen, Rohit Pillai, Yu Cheng, Luowei Zhou, Xin Eric Wang, William Yang Wang, et al. VALUE: A Multi-Task Benchmark for Video-and-Language Understanding Evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[62] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic Evaluation of Language Models. *arXiv preprint arXiv:2211.09110*, 2022.

[63] Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu,

Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online, November 2020. Association for Computational Linguistics.

[64] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*, 2019.

[65] Ekaterina Lutikova. K voprosu o kategorial'nom statuse imennykh grup v russkom yazyke. *Moscow University Philology Bulletin*, 2010.

[66] Zhiyi Ma, Kawin Ethayarajh, Tristan Thrush, Somya Jain, Ledell Wu, Robin Jia, Christopher Potts, Adina Williams, and Douwe Kiela. Dynaboard: An Evaluation-As-A-Service Platform for Holistic Next-Generation Benchmarking. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 10351–10367. Curran Associates, Inc., 2021.

[67] Brian W. Matthews. Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochimica et biophysica acta*, 405 2:442–51, 1975.

[68] Kris McGuffie and Alex Newhouse. The Radicalization Risks of GPT-3 and Advanced Neural Language Models. *arXiv preprint arXiv:2009.06807*, 2020.

[69] Michail Melnichenko and Natalia Tyshkevich. Prozhito from Manuscript to Corpus. *ISTORIYA*, 8(7 (61)), 2017.

[70] Vladislav Mikhailov, Tatiana Shamardina, Max Ryabinin, Alena Pestova, Ivan Smurov, and Ekaterina Artemova. RuCoLA: Russian corpus of linguistic acceptability. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5207–5227, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[71] Swaroop Mishra and Anjana Arunkumar. How Robust are Model Rankings: A Leaderboard Customization Approach for Equitable Evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13561–13569, 2021.

[72] Olga Mitrenina, Evgeniya Romanova, and Natalia Slioussar. *Vvedeniye v generativnuyu grammatiku*. Limited Liability Company "LIBROCOM", 2017.

[73] Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language*

Processing (Volume 1: Long Papers), pages 5356–5371, Online, August 2021. Association for Computational Linguistics.

[74] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online, November 2020. Association for Computational Linguistics.

[75] Christina Nießl, Moritz Herrmann, Chiara Wiedemann, Giuseppe Casalicchio, and Anne-Laure Boulesteix. Over-optimism in Benchmark Studies and the Multiplicity of Design and Analysis Options when Interpreting Their Results. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(2):e1441, 2022.

[76] Hannu Nurmi. Voting procedures: A summary analysis. *British Journal of Political Science*, 13(2):181–208, 1983.

[77] Elena Paducheva. *Dinamicheskiye modeli v semantike leksiki*. Languages of Slavonic culture, 2004.

[78] Elena Paducheva. *Semanticheskiye issledovaniya: Semantika vremeni i vida v russkom yazyke*. Languages of Slavonic culture, second edition, 2010.

[79] Elena Paducheva. *Russkoye otritsatel'noye predlozheniye*. Languages of Slavonic culture, 2013.

[80] Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Ji Yoon Han, Jangwon Park, Chisung Song, Junseong Kim, Youngsook Song, Taehwan Oh, et al. KLUE: Korean Language Understanding Evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[81] Alicia Parrish, William Huang, Omar Agha, Soo-Hwan Lee, Nikita Nangia, Alexia Warstadt, Karmanya Aggarwal, Emily Allaway, Tal Linzen, and Samuel R. Bowman. Does putting a linguist in the loop improve NLU data collection? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4886–4901, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[82] Adam Pauls and Dan Klein. Large-scale syntactic language modeling with treelets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 959–968, 2012.

[83] Maxime Peyrard, Teresa Botschen, and Iryna Gurevych. Learning to score system summaries for better content selection evaluation. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 74–84, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[84] Irina Proskurina, Irina Piontkovskaya, and Ekaterina Artemova. Can BERT Eat RuCoLA? Topological Data Analysis to Explain. *arXiv preprint arXiv:2304.01680*, 2023.

[85] Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[86] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language Models are Unsupervised Multitask Learners. 2019.

[87] Mark Rofin, Vladislav Mikhailov, Mikhail Florinsky, Andrey Kravchenko, Tatiana Shavrina, Elena Tutubalina, Daniel Karabekyan, and Ekaterina Artemova. Vote'n'rank: Revision of benchmarking with social choice theory. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 670–686, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[88] Anna Rogers. How the Transformers Broke NLP Leaderboards. `https://hackingsemantics.xyz/2019/leaderboards`, 2019.

[89] Anna Rogers, Matt Gardner, and Isabelle Augenstein. QA Dataset Explosion: A Taxonomy of NLP Resources for Question Answering and Reading Comprehension. *ACM Computing Surveys*, 55(10):1–45, 2023.

[90] Sebastian Ruder. Why You Should Do NLP Beyond English. `http://ruder.io/nlp-beyond-english`, 2020.

[91] Sebastian Ruder. Challenges and Opportunities in NLP Benchmarking. `http://ruder.io/nlp-benchmarking`, 2021.

[92] Sebastian Ruder. The State of Multilingual AI. `http://ruder.io/state-of-multilingual-ai/`, 2022.

[93] Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[94] Piotr Rybak, Robert Mroczkowski, Janusz Tracz, and Ireneusz Gawlik. KLEJ: Comprehensive benchmark for Polish language understanding. In *Proceedings of the 58th Annual*

Meeting of the Association for Computational Linguistics, pages 1191–1201, Online, July 2020. Association for Computational Linguistics.

[95] Andrey Sakhovskiy, Alexandra Izhevskaya, Alena Pestova, Elena Tutubalina, Valentin Malykh, Ivan Smurov, and Ekaterina Artemova. RuSimpleSentEval-2021 Shared Task: Evaluating Sentence Simplification for Russian. In *Proceedings of the International Conference "Dialogue"*, pages 607–617, 2021.

[96] Gerard Salton and C. S. Yang. On the specification of term values in automatic indexing. *Journal of Documentation*, 29:351–372, 1973.

[97] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[98] Ketki Savle, Wlodek Zadrozny, and Minwoo Lee. Topological data analysis for discourse semantics? In *Proceedings of the 13th International Conference on Computational Semantics - Student Papers*, pages 34–43, Gothenburg, Sweden, May 2019. Association for Computational Linguistics.

[99] Carson T. Schütze. *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. University of Chicago Press, 1996.

[100] Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online, April 2021. Association for Computational Linguistics.

[101] Haitham Seelawi, Ibraheem Tuffaha, Mahmoud Gzawi, Wael Farhan, Bashar Talafha, Riham Badawi, Zyad Sober, Oday Al-Dweik, Abed Alhakim Freihat, and Hussein Al-Natsheh. ALUE: Arabic language understanding evaluation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 173–184, Kyiv, Ukraine (Virtual), April 2021. Association for Computational Linguistics.

[102] Olga Seliverstova. *Trudy po semantike*. Languages of Slavonic culture, 2004.

[103] Tatiana Shamardina, Vladislav Mikhailov, Daniil Chernianskii, Alena Fenogenova, Marat Saidov, Anastasiya Valeeva, Tatiana Shavrina, Ivan Smurov, Elena Tutubalina, and Ekaterina Artemova. Findings of the RuATD Shared Task 2022 on Artificial Text Detection in Russian. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2022"*, 2022.

[104] Tatiana Shavrina, Anton Emelyanov, Alena Fenogenova, Vadim Fomin, Vladislav Mikhailov, Andrey Evlampiev, Valentin Malykh, Vladimir Larin, Alex Natekin, Alek-

sandr Vatulin, Peter Romov, Daniil Anastasiev, Nikolai Zinov, and Andrey Chertok. Humans keep it one hundred: an overview of AI journey. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2276–2284, Marseille, France, May 2020. European Language Resources Association.

[105] Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. RussianSuperGLUE: A Russian language understanding evaluation benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4717–4726, Online, November 2020. Association for Computational Linguistics.

[106] Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. Release Strategies and the Social Impacts of Language Models, 2019.

[107] Anirudh Som, Hongjun Choi, Karthikeyan Natesan Ramamurthy, and Matthew P. Buman. Pi-net: A Deep Learning Approach to Extract Topological Persistence Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

[108] Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual Translation with Extensible Multilingual Pretraining and Finetuning, 2020.

[109] Yakov Testelets. *Vvedeniye v obschiy sintaksis*. Russian State University for the Humanities, 2001.

[110] Jörg Tiedemann and Santhosh Thottingal. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal, November 2020. European Association for Machine Translation.

[111] Daniela Trotta, Raffaele Guarasci, Elisa Leonardelli, and Sara Tonelli. Monolingual and cross-lingual acceptability judgments with the Italian CoLA corpus. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2929–2940, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[112] Eduard Tulchinskii, Kristian Kuznetsov, Laida Kushnareva, Daniil Cherniavskii, Serguei Barannikov, Irina Piontkovskaya, Sergey Nikolenko, and Evgeny Burnaev. Topological Data Analysis for Speech Processing. *arXiv preprint arXiv:2211.17223*, 2022.

[113] Alan M Turing and J Haugeland. Computing machinery and intelligence. *The Turing Test: Verbal Behavior as the Hallmark of Intelligence*, pages 29–56, 1950.

[114] Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. Authorship attribution for neural text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8384–8395, Online, November 2020. Association for Computational Linguistics.

[115] Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. TURINGBENCH: A benchmark environment for Turing test in the age of neural text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2001–2016, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[116] Gorka Urbizu, Iñaki San Vicente, Xabier Saralegi, Rodrigo Agerri, and Aitor Soroa. BasqueGLUE: A natural language understanding benchmark for Basque. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1603–1612, Marseille, France, June 2022. European Language Resources Association.

[117] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. *Advances in Neural Information Processing Systems*, 30, 2017.

[118] Elena Volodina, Yousuf Ali Mohammed, and Julia Klezl. DaLAJ – a dataset for linguistic acceptability judgments for Swedish. In *Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning*, pages 28–37, Online, May 2021. LiU Electronic Press.

[119] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. *Advances in Neural Information Processing Systems*, 32, 2019.

[120] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[121] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *International Conference on Learning Representations*, 2019.

[122] Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. Adversarial GLUE: A Multi-Task Benchmark for

Robustness Evaluation of Language Models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[123] W Wang, N Yang, F Wei, B Chang, and M Zhou. R-NET: Machine Reading Comprehension with Self-matching Networks. *Natural Lang. Comput. Group, Microsoft Res. Asia, Beijing, China, Tech. Rep*, 5, 2017.

[124] Alex Warstadt and Samuel R Bowman. Linguistic Analysis of Pretrained Sentence Encoders with Acceptability Judgments. *arXiv preprint arXiv:1901.03438*, 2019.

[125] Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392, 2020.

[126] Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 2019.

[127] Geoffrey I Webb. MultiBoosting: A Technique for Combining Boosting and Wagging. *Machine learning*, 40(2):159–196, 2000.

[128] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Ethical and Social Risks of Harm from Language Models, 2021.

[129] Mattthew E Werenski, Ruijie Jiang, Abiy Tasissa, Shuchin Aeron, and James M Murphy. Measure Estimation in the Barycentric Coding Model. In *International Conference on Machine Learning*, pages 23781–23803. PMLR, 2022.

[130] Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, and Ayu Purwarianti. IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 843–857, Suzhou, China, December 2020. Association for Computational Linguistics.

[131] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[132] Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. CLiMP: A benchmark for Chinese language model evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2784–2790, Online, April 2021. Association for Computational Linguistics.

[133] Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[134] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June 2021. Association for Computational Linguistics.

[135] Fan Yin, Quanyu Long, Tao Meng, and Kai-Wei Chang. On the robustness of language encoders against grammatical errors. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3386–3403, Online, July 2020. Association for Computational Linguistics.

[136] Kuratov Yuri and Arkhipov Mikhail. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2019"*, 2019.

[137] Aleš Žagar and Marko Robnik-Šikonja. Slovene SuperGLUE benchmark: Translation and evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2058–2065, Marseille, France, June 2022. European Language Resources Association.

[138] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. *Advances in neural information processing systems*, 32, 2019.

[139] Changchang Zeng, Shaobo Li, Qin Li, Jie Hu, and Jianjun Hu. A Survey on Machine Reading Comprehension – Tasks, Evaluation Metrics and Benchmark Datasets. *Applied Sciences*, 10(21):7640, 2020.

[140] Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. ReCoRD: Bridging the Gap Between Human and Machine Commonsense Reading Comprehension. *arXiv preprint arXiv:1810.12885*, 2018.

[141] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[142] Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404, Online, August 2021. Association for Computational Linguistics.

[143] Xiyou Zhou, Zhiyu Chen, Xiaoyong Jin, and William Yang Wang. HULK: An energy efficiency benchmark platform for responsible natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 329–336, Online, April 2021. Association for Computational Linguistics.