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

**Tatiana Shamardina**[1*], **Vladislav Mikhailov**[2*], **Daniil Chernianskii**[3,4],
**Alena Fenogenova**[2], **Marat Saidov**[6], **Anastasiya Valeeva**[7],
**Tatiana Shavrina**[2,3], **Ivan Smurov**[1,7], **Elena Tutubalina**[5,6,8], **Ekaterina Artemova**[6]

[1]ABBYY, [2]SberDevices, [3]AIRI, [4]Skolkovo Institute of Science and Technology, [5]Sber AI,
[6]HSE University, [7]Moscow Institute of Physics and Technology,[8]Kazan Federal University

**Correspondence**: tatiana.shamardina@abbyy.com

**Abstract**

We present the shared task on artificial text detection in Russian, which is organized as a part of the Dialogue Evaluation initiative, held in 2022. The shared task dataset includes texts from 14 text generators, i.e., one human writer and 13 text generative models fine-tuned for one or more of the following generation tasks: machine translation, paraphrase generation, text summarization, text simplification. We also consider back-translation and zero-shot generation approaches. The human-written texts are collected from publicly available resources across multiple domains.

The shared task consists of two sub-tasks: (i) to determine if a given text is automatically generated or written by a human; (ii) to identify the author of a given text. The first task is framed as a binary classification problem. The second task is a multi-class classification problem. We provide count-based and BERT-based baselines, along with the human evaluation on the first sub-task. A total of 30 and 8 systems have been submitted to the binary and multi-class sub-tasks, correspondingly. Most teams outperform the baselines by a wide margin. We publicly release our codebase, human evaluation results, and other materials in our GitHub repository.

**Keywords:** artificial text detection, natural language generation, shared task, neural authorship attribution, transformers

# RuATD-2022: Соревнование по автоматическому распознаванию сгенерированных текстов

**Татьяна Шамардина**[1*], **Владислав Михайлов**[2*], **Даниил Чернявский**[3,4],
**Алена Феногенова**[2], **Марат Саидов**[6], **Анастасия Валеева**[7],
**Татьяна Шаврина**[2,3], **Иван Смуров**[1,7], **Елена Тутубалина**[5,6,8], **Екатерина Артемова**[6]

[1]ABBYY, [2]SberDevices, [3]AIRI, [4]Сколтех, [5]Sber AI,
[6]НИУ ВШЭ, [7]МФТИ,[8]КФУ

**Для связи**: tatiana.shamardina@abbyy.com

**Аннотация**

Данная статья представляет собой отчет организаторов соревнования RuATD-2022, посвященного автоматическому распознаванию сгененированных текстов на материале русского языка. Соревнование RuATD-2022 проходило в рамках кампании Dialogue Evaluation в 2022 году. Набор данных, использованный в соревновании, частично составлен автоматически с использованием моделей генерации текстов. Мы использовали модели, обученные решать различные задачи генерации текстов: машинного перевода, генерация парафразов, автоматического реферирования и упрощения предложений. Мы также рассматриваем популярные постановки задач, такие как обратный перевод и zero-shot генерация. Вторая часть набора данных – тексты, написанные людьми – собрана из открытых источников, относящихся к ряду предметных областей.

Участникам соревнования предлагается решить две задачи: (i) определить, был ли данный текст написан человеком или сгенерирован моделью (бинарная классификация), или (ii) определить автора текста (мультиклассовая классификация). В рамках соревнования мы предоставляем базовые решения в стандартной постановке задачи классификации на основе счетных признаков (TF-IDF) и модели архитектуры BERT. Кроме того, мы проводим оценку решения первой задачи разметчиками на краудсорсинговой платформе (human baseline). В общей сложности, соревнование привлекло внимание 38 решений: 30 для первой постановки задачи и 8 – для второй. Большая

---

*Equal contribution.

Shamardina T., Mikhailov V., Chernianskii D., Fenogenova A., Saidov M., Valeeva A., Shavrina T., Smurov I., Tutubalina E., Artemova E.

часть участников преодолела уровень базовых решений и уровень разметчиков. Используемая кодовая база, результаты оценки на краудсорсинговой платформе и другие материалы соревнования доступны в публичном GitHub репозитории соревнования.

**Ключевые слова:** распознавание сгенерированных текстов, генерация текстов, соревнование, автоматическое определение автора текста, нейронные сети

## 1 Introduction

Modern text generative models (TGMs) have demonstrated impressive results in generating texts close to the human level in terms of fluency, coherence, and grammar (Keskar et al., 2019; Zellers et al., 2019; Brown et al., 2020; Rae et al., 2021). However, the misuse potential of TGMs increases with their capabilities to generate more human-like texts. Malicious users can deploy TGMs for spreading propaganda and fake news (Zellers et al., 2019; Uchendu et al., 2020; McGuffie and Newhouse, 2020), augmenting fake product reviews (Adelani et al., 2020), and facilitating fraud, scams, and other targeted manipulation (Weidinger et al., 2021). The increasing difficulty for laypeople and users to discriminate machine-generated texts from human-written ones facilitates the spread of such misuse (Karpinska et al., 2021; Uchendu et al., 2021). This motivates the *artificial text detection* task (Jawahar et al., 2020), a fast-growing niche field aimed at mitigating the misuse of TGMs.

The Russian Artificial Text Detection (RuATD) shared task explores the problem of artificial text detection in Russian. Unlike existing datasets for English, our approach includes a range of task-specific TGMs, that is, models fine-tuned for common text generation tasks at the sentence- and document-level. On the one hand, such a setting challenges the participants and crowd-sourced annotators. On the other hand, it also enables many research and development purposes, such as training and benchmarking artificial text detectors, warning users about potentially fake content on social media and news platforms, filtering corpora augmented with TGMs, exploring detectors' robustness w.r.t. TGMs' architecture, size, downstream task, or domain. The shared task dataset consists of publicly available texts across multiple domains and texts generated by various monolingual and cross-lingual TGMs. The setup includes two sub-tasks: (i) to determine if a given text is automatically generated or written by a human (binary classification), and (ii) to identify the author of a given text (multi-class classification).

The main contributions of this paper are the following:

1. We propose a diverse automatic text detection corpus in Russian, the first of its kind (§2.2);
2. We model two competition sub-tasks (§4.1) after the traditional concepts of "Turing test" and authorship attribution for neural text generation models (Uchendu et al., 2021). We establish two count-based and neural-based baseline solutions (§4.4) and the human evaluation on the binary classification problem (§4.6);
3. We conduct an extensive analysis of the received submissions for both sub-tasks (§5) and discuss potential research directions (§6);
4. We set up the shared task environment, which remains open for the community submissions to facilitate future research in the area (§4.3).

## 2 Dataset

### 2.1 Text Generation

The corpus includes texts from 14 text generators, i.e., one human writer and 13 monolingual/multilingual TGMs varying in their number of parameters, architecture choices, and pre-training objectives. Each model is fine-tuned for one or more of the following text generation tasks: MT, paraphrase generation, text simplification, and text summarization. We also consider back-translation and zero-shot generation approaches. Figure 1 outlines the dataset creation pipeline. Text generation hyperparameters for each model are presented in 1. Note that we mostly use the default hyperparameters under the considered libraries.

**Human** We collect human-written texts from publicly available resources among six domains (see Section 2.2 for more details). Gold standard references from task-specific datasets are also used as human texts, since they are generally written and/or validated by crowd-source annotators (Artetxe and
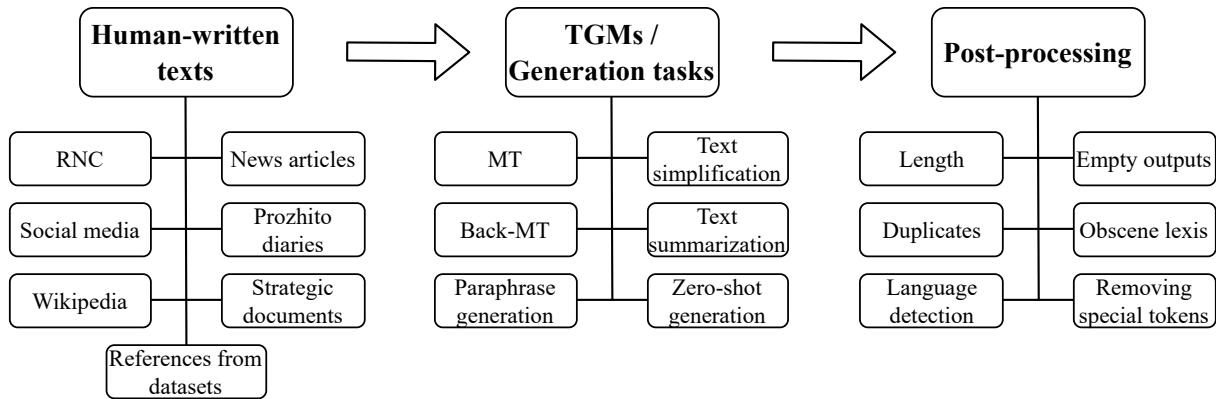
Figure 1: Data collection, text generation, and post-processing procedure.

| Task | Model | Hyperparameters |
|---|---|---|
| **Machine Translation** | OPUS-MT<br>M-BART50<br>M2M-100 | BEAM SEARCH; num beams=5 |
| **Back Translation** | OPUS-MT<br>M-BART50<br>M2M-100 | BEAM SEARCH; num beams=5 |
| **Paraphrase Generation** | mT5-Small, mT5-Large<br>ruT5-Base-Multitask<br>ruGPT2-Large, ruGPT3-Large | TOP-P SAMPLING; p=0.9 |
| **Simplification** | mT5-Large, ruT5-Large<br>ruGPT3-Small, ruGPT3-Medium, ruGPT3-Large | TOP-P SAMPLING; p=0.9 |
| **Summarization** | M-BART<br>ruT5-Base | BEAM SEARCH; num beams=5 |
| **Zero-shot Generation** | ruGPT3-Small, ruGPT3-Medium<br>ruGPT3-Large | TOP-P SAMPLING; p=0.95;<br>max length=90 percentile of length distribution by domain |

Table 1: A brief description of the text generation hyperparameters and decoding strategies by text generation task.

Schwenk, 2019; Schwenk et al., 2021; Scialom et al., 2020; Hasan et al., 2021). The human texts serve as the input to the TGMs.

**MT & Back-translation** We use three MT models via the `EasyNMT` framework[1]: OPUS-MT (Tiedemann and Thottingal, 2020), M-BART50 (Tang et al., 2020), and M2M-100 (Fan et al., 2021). We use subsets of the Tatoeba (Artetxe and Schwenk, 2019) and WikiMatrix (Schwenk et al., 2021) datasets to obtain translations among three language pairs: English-Russian, French-Russian, and Spanish-Russian. In the back-translation setting, the input sentence is translated into one of the target languages, and then back into Russian.

**Paraphrase Generation** Paraphrases are generated with models available under the `russian-paraphrasers` library (Fenogenova, 2021a): ruGPT2-Large[2], ruT5-Base-Multitask[3], and mT5 (Xue et al., 2021) of Small and Large versions.

**Text Simplification** We fine-tune ruGPT3-Small[4], ruGPT3-Medium[5], ruGPT3-Large[6], mT5-Large, and ruT5-Large[7] for text simplification on a filtered version of the RuSimpleSentEval-2022 dataset (Sak-

---

[1] `github.com/UKPLab/EasyNMT`

[2] `hf.co/sberbank-ai/rugpt2large`

[3] `hf.co/cointegrated/rut5-base-multitask`

[4] `hf.co/sberbank-ai/rugpt3small`

[5] `hf.co/sberbank-ai/rugpt3medium`

[6] `hf.co/sberbank-ai/rugpt3large`

[7] `hf.co/sberbank-ai/rugt5-large`

hovskiy et al., 2021; Fenogenova, 2021b). Fine-tuning of each model is run for 4 epochs with the batch size of 4, learning rate of $10^{-5}$, and weight decay of $10^{-2}$.

**Text Summarization** We use two abstractive summarization models fine-tuned on the Gazeta dataset (Gusev, 2020): ruT5-base[8] and M-BART[9].

**Zero-shot Generation** We generate texts in a zero-shot manner by prompting the model and specifying the maximum number of generated tokens. The models include ruGPT3-Small, ruGPT3-Medium, ruGPT3-Large.

## 2.2 Data

Pre-training corpora of TGMs can cover multiple versatile domains (Liu et al., 2020), which prompt their abilities to generate texts with specific lexical, syntactic, discourse and stylistic properties. Despite this, the ATD task is generally explored w.r.t. only one particular domain, e.g., product reviews (Adelani et al., 2020), social media posts (Fagni et al., 2021), or news (Uchendu et al., 2021). Such setting limits the scope of evaluation of artificial text detectors. A few studies show that performance of modern detectors can vary drastically across domains (Bakhtin et al., 2019; Kushnareva et al., 2021), which stimulates the development of more generalizable and robust methods (Jawahar et al., 2020).

This paper aims at providing a diverse shared task data, taking into account the current limitations in the niche ATD field, and the diversity of TGMs widely used in the industry and NLP research for Russian. To this end, we consider domains which represent normative Russian, as well as general domain texts, social media posts, texts of different historical periods, bureaucratic texts with complex discourse structure and embedded named entities, and other domains included in the task-specific datasets, such as subtitles and web-texts. Recall that aside from linguistic and stylometric properties, texts differ in their length (e.g., sentence-level vs. document-level), and specifics attributable to the downstream tasks. We now list domains of texts that are fed into the previously described TGMs.

**Russian National Corpus** We use the diachronic sub-corpora of the Russian National Corpus[10] (RNC), which covers three historical periods of the society and the Modern Russian language ("pre-Soviet", "Soviet", and "post-Soviet").

**Social Media** We parse texts from multiple social media platforms that are marked with certain hashtags, such as dates, months, seasons, holidays, the names of large cities in Russia, etc. These texts are typically short, written in informal style and may contain emojis and obscene lexis.

**Wikipedia** We select the top-100 most viewed Russian Wikipedia pages spanning the period of 2016-2021 according to the PageViews[11] statistics.

**News Articles** The news segment covers different news sources in the Taiga corpus (Shavrina and Shapovalova, 2017) and the `corus` library[12], including but not limited to Lenta, KP, Interfax, Izvestia, Gazeta. We additionally parse more recent news articles to prevent potential data leakage and cheating.

**Prozhito Diaries** Prozhito is a corpus of digitilized personal diaries, written during the 20th century (Melnichenko and Tyshkevich, 2017).

**Strategic Documents** are produced by the Ministry of Economic Development of the Russian Federation. The documents are written in bureaucratic style, rich in embedded entities, and have complex syntactic and discourse structure. This dataset has been previously used in the RuREBus shared task (Ivanin et al., 2020).

## 2.3 Post-processing

Each generated text undergoes a post-processing procedure based on a combination of language processing tools and heuristics. First, we discard duplicates, copied inputs, empty outputs, and remove special tokens from the generated texts (e.g., `<s>`, `</s>`, `<pad>`, etc.). Next, we empirically define length

---

[8] `hf.co/IlyaGusev/rut5-base-sum-gazeta`
[9] `hf.co/IlyaGusev/mbart-ru-sum-gazeta`
[10] `ruscorpora.ru`
[11] `https://pageviews.wmcloud.org/`
[12] `github.com/natasha/corus`

| Task | Text Generator | Domain | Num. samples | Num. tokens | IPM |
|---|---|---|---|---|---|
| **Machine Translation** | Human<br>OPUS-MT<br>M-BART50<br>M2M-100 | Tatoeba<br>WikiMatrix | 35860 | 11.5 | 0.89 |
| **Back Translation** | Human<br>OPUS-MT<br>M-BART50<br>M2M-100 | Strategic documents<br>News<br>Prozhito<br>RNC<br>Wikipedia<br>Tatoeba<br>WikiMatrix | 35588 | 12.9 | 0.88 |
| **Paraphrase Generation** | Human<br>mT5-Small<br>mT5-Large<br>ruT5-Base-Multitask<br>ruGPT2-Large<br>ruGPT3-Large | Strategic documents<br>News<br>Prozhito<br>RNC<br>Wikipedia<br>Social media | 44298 | 13.0 | 0.85 |
| **Simplification** | Human<br>mT5-Large<br>ruT5-Large<br>ruGPT3-Small<br>ruGPT3-Medium<br>ruGPT3-Large | Strategic documents<br>News<br>Prozhito<br>RNC<br>Wikipedia<br>Social media | 44700 | 18.3 | 0.86 |
| **Summarization** | Human<br>M-BART<br>ruT5-Base | Strategic documents<br>News<br>Prozhito<br>RNC<br>Wikipedia | 17164 | 33.5 | 0.86 |
| **Zero-shot Generation** | Human<br>ruGPT3-Small<br>ruGPT3-Medium<br>ruGPT3-Large | Strategic documents<br>News<br>Prozhito<br>RNC<br>Wikipedia | 37499 | 141.5 | 0.85 |

Table 2: Text generators, domains and the final number of samples per task. The number of human-written texts is same as machine-generated texts.

intervals for each generation task based on a manual analysis of length distributions in `razdel`[13] tokens. The texts are filtered by the following token ranges: 5-to-25 (**MT, Back-translation, Paraphrase Generation**), 10-to-30 (**Text Simplification**), 15-to-60 (**Text Summarization**), and 85-to-400 (**Zero-shot Generation**). We additionally discard the social media texts containing obscene lexis according to the corpus of Russian obscene words[14], and keep the MT/Back-translation texts classified as Russian with the confidence of more than $0.9$ (`langdetect`[15]).

## 3 Dataset Statistics

This section describes various count-based statistics of our dataset for human-written and machine-generated texts.

**General Statistics** Table 2 shows general dataset statistics w.r.t. text generation task, text generator, and domain. On average, there are $37.9$ tokens in each text, with variations depending on the task. We estimate the frequency of each text according to the Russian National Corpus (RNC)[16]. It is computed as the number of frequently used tokens (i.e., the number of instances per million, that is, IPM in RNC is higher than 1) divided by the number of tokens in a sentence. The average IPM is $0.86$ for the human-written texts and $0.87$ for the machine-generated ones.

**Diversity Metrics** We estimate the diversity of the texts in terms of their k-gram statistics and lexical richness. We calculate two diversity metrics upon $k$-gram statistics: Dist-$k$ (Li et al., 2016) and Ent-

---

[13]`github.com/natasha/razdel`
[14]`github.com/odaykhovskaya/obscene-words`
[15]`github.com/fedelopez77/langdetect`
[16]`ruscorpora.ru/new/en`

|  | k | H | M |
|---|---|---|---|
| **Dist-k** | 1 | 0.35 | **0.40** |
| | 2 | 0.75 | **0.76** |
| | 4 | 0.74 | **0.77** |
| **Ent-k** | 1 | **7.86** | 7.30 |
| | 2 | **9.99** | 9.03 |
| | 4 | **10.16** | 9.19 |

Table 3: Dist-$k$ and Ent-$k$ diversity measures by the target level. **H**=Human-written texts; **M**=Machine-generated texts.

|  | Back-MT | | MT | | Zero-shot Gen. | | Paraphrase Gen. | | Simplification | | Summarization | | Overall | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  | H | M | H | M | H | M | H | M | H | M | H | M | H | M |
| **Words** | 10.04 | **10.66** | **9.65** | 8.84 | 106.05 | **112.48** | 10.21 | **11.64** | 14.48 | **14.70** | 24.46 | **30.34** | 28.82 | **30.72** |
| **Terms** | 9.70 | **10.07** | **9.33** | 8.47 | 73.52 | **95.40** | 9.87 | **11.33** | 13.72 | 13.55 | 22.32 | **26.16** | 22.64 | **26.95** |
| **TTR** | 0.95 | 0.95 | **0.98** | 0.97 | 0.70 | **0.86** | 0.96 | **0.97** | 0.95 | 0.93 | **0.93** | 0.87 | 0.91 | **0.93** |
| **CTTR** | 2.09 | **2.13** | **2.08** | 1.97 | 4.91 | **6.20** | 2.13 | **2.31** | **2.52** | 2.47 | 3.14 | **3.33** | 2.76 | **3.01** |

Table 4: Lexical richness metrics per text generation task.

$k$ (Zhang et al., 2018). Dist-$k$ is the total number of k-grams divided by the number of tokens in the text set. Ent-$k$ is an entropy metric that weights each k-gram so infrequent k-grams are penalized and contribute less to diversity. We compute the diversity scores for texts grouped by label and report them for $k \in \{1, 2, 4\}$ in Table 3.

To measure the lexical diversity[17] of the texts in our dataset, we calculate four types of metrics: word count, terms count, type-token ratio (TTR), and corrected type-token ratio (CTTR). Type-token ratio is computed as $t/w$ and corrected type-token ratio is computed as $t/\sqrt{2*w}$, where $t$ is the number of unique terms/vocabulary, and $w$ is the total number of words.

We can see that the ratio of the diversity measures between the natural and artificial texts depends on the task, which is explained by the very task formulation. At the same time, artificial texts may include non-existent words, degenerated textual segments, or rare words, which can be attributed to more significant lexical richness metrics overall.

## 4 Setup

### 4.1 Tasks

The RuATD Shared task features two sub-tasks:

I. Determine if a given text is automatically generated or written by a human. This sub-task is framed as a binary classification problem with two labels: **H** (human) and **M** (machine).

II. Identify the author of a given text. This sub-task is modeled after the traditional problem of authorship attribution (Coyotl-Morales et al., 2006), particularly in the context of neural models (Uchendu et al., 2020). It is a multi-class classification problem with 14 target classes – a human writer and 13 TGMs.

**Evaluation** Each sub-task uses the accuracy score, a standard metric for classification with balanced classes, as the official evaluation metric.

### 4.2 Dataset Splits

We split the dataset into four sets in the 60/10/15/15 proportion ratio: train (130k), development (21k), public test (32k), and private test (32k). Each set is balanced by the number of target classes, text generator, text generation task, and domain[18]. These sets are used for both sub-tasks, with only the target

---

[17]Lexical richness

[18]The number of human-written texts is equal to the number of machine-generated texts for each domain and text generation task.

Определите: текст написан человеком или ИИ?

Экономическая эффективность реализации подпрограммы зависит от степени достигаемого конечных результатов.
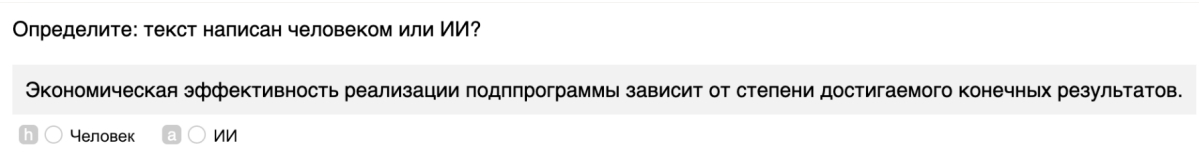
h ○ Человек    a ○ ИИ

Figure 2: An example of the Toloka interface for the human evaluation setup.

classes changed, i.e., the **M** label is broken into 13 TGMs' names in the multi-class sub-task.

### 4.3  Kaggle Setup

We use the Kaggle competition platform to run the shared task. The sub-tasks are set as separate competitions and leaderboards:

I. The binary sub-task is hosted under this link;

II. The multi-class sub-task is hosted under this link.

The participants are allowed to take part solely or in teams in both sub-tasks. The shared task comprises two stages: **public** and **private** testing. The first stage provides access to the public test set and leaderboard, allowing the participants to develop and improve their submissions during the competition. The second stage defines the final leaderboard ranking on the private test set, scoring up to three submissions selected by the participants. Otherwise, the Kaggle platform automatically selects the three best submissions based on the participants' public test scores. Participants are allowed to use any additional materials and pre-trained models, except for direct markup of the test set and search on the Internet.

### 4.4  Baseline

We provide the participants with two open-source baseline solutions: count-based (TF-IDF baseline) and BERT-based (BERT baseline) (Devlin et al., 2019). TF-IDF baseline is based on TF-IDF features coupled with the SVD dimensionality reduction and a Logistic Regression classifier. The TF-IDF has 50k features, further reduced to 5000 by SVD. The BERT baseline follows the default fine-tuning and evaluation procedure for the classification task under the HuggingFace transformers framework (Wolf et al., 2020).

### 4.5  Peer Review

Each participant is asked to publicly release their solutions and peer review other participants' submissions. This step allows for a fair evaluation, eliminating the risks of potential cheating, such as solving the sub-tasks via a web search or other heuristics. After analyzing the assigned submission, the peer-reviewer should answer two questions in the Google form and provide comments, if any:

• Does the submission use a web search?

• Does the submission violate any other rules[19] of the shared task?

### 4.6  Human Baseline

We conduct a human evaluation on the binary classification problem using stratified sub-samples from the public and private test sets. Each subset of 2.5k samples is balanced by the number of target classes, text generator, text generation task, and domain. We report the human baseline results during both public and private testing stages. The evaluation is run via the Toloka platform. The annotation setup follows the conventional crowd-sourcing guidelines for the ATD task and accounts for methodological limitations discussed in (Ippolito et al., 2020; Clark et al., 2021; Karpinska et al., 2021). We provide a full annotation instruction in Figure 1 (see Appendix A), and an example of the Toloka interface in Figure 2.

---

[19]The shared task rules are provided in the GitHub repository.

| Binary sub-task | | Multi-class sub-task | |
|---|---|---|---|
| **Team** | **Accuracy** | **Team** | **Accuracy** |
| MSU ✓ | 0.82995 | Posokhov Pavel ✓ | 0.65035 |
| Igor | 0.82725 | Yixuan Weng ✓ | 0.64731 |
| Orzhan ✓ | 0.82629 | Orzhan ✓ | 0.64573 |
| mariananieva ✓ | 0.82427 | MSU ✓ | 0.62856 |
| Ivan Zakharov | 0.82294 | BERT baseline | 0.59813 |
| Yixuan Weng ✓ | 0.81767 | Nikita Selin | 0.58967 |
| ilya koziev | 0.81699 | Victor Krasilnikov | 0.55012 |
| miso soup ✓ | 0.81178 | Petr Grigoriev ✓ | 0.45814 |
| Eduard Belov | 0.80862 | TF-IDF baseline | 0.44280 |
| Posokhov Pavel ✓ | 0.80630 | Anastasiya Shabaeva | 0.05411 |
| Kirill Apanasovich | 0.80308 | | |
| Tumanov Alexander | 0.79778 | | |
| BERT baseline | 0.79666 | | |
| Elizaveta Nosova | 0.79595 | | |
| mipatov | 0.78591 | | |
| akstar | 0.78442 | | |
| Nikita Selin | 0.78228 | | |
| David Avagyan | 0.77869 | | |
| Mikhail Yumanov | 0.77181 | | |
| Gregory Kuznetsov | 0.75237 | | |
| Anastasiya Shabaeva | 0.75178 | | |
| Shershunya | 0.74534 | | |
| Ekaterina Kostrykina | 0.74326 | | |
| Victor Krasilnikov | 0.74091 | | |
| Alena | 0.73589 | | |
| Alexander Tesemnikov | 0.73204 | | |
| Lera Lelik | 0.72727 | | |
| Dmitriy Vahrushev | 0.71559 | | |
| Human baseline | 0.66666 | | |
| Molostvov Pavel | 0.68543 | | |
| Mental Sky | 0.65326 | | |
| Petr Grigoriev ✓ | 0.64232 | | |
| TF-IDF baseline | 0.64223 | | |

Table 5: The official shared task results sorted in the descending order. Left: the binary leaderboard; Right: the multi-class leaderboard. Baseline submissions are colored in grey. ✓ stands for peer-reviewed submissions.

We grant access to the human evaluation project to only top-70% annotators according to the in-house Toloka rating system. Each annotator must first finish the training task by completing at least 80% of samples correct to get onto the main annotation task. We use the dynamic overlap of 3-to-5 annotators per sample. We discard votes from those annotators whose quality rate on the control tasks is less than 50%. We also filter out votes with the response time of less than 15 seconds per annotation task page (5 samples). The resulting vote is aggregated as the majority vote label.

## 5   Results and Analysis

We report the official shared task results of the private testing stage in Table 5. Only peer-reviewed submissions (marked with ✓ in the Table) are considered official.

As one can notice from Table, top-4 systems in the binary classification task have been peer-reviewed. In the multi-class setting, all four top-4 systems have been also peer-reviewed.

The results demonstrate that state-of-the-art classification models can be relatively successful in distinguishing human-written texts from machine-generated ones for the Russian language and determining the exact model used for generation for the latter class. However, one can quickly notice a rather stark contrast between the best scores obtained on the RuATD test set in binary setup (0.830 accuracy for *MSU*, the top-system in binary classification task) and scores obtained for a similar setup in English (0.970 accuracy; see (Uchendu et al., 2020) for reference). We attribute this contrast not to the difference in the languages but mainly to the nature of texts: in the English setup, an average text length is 432 words (compared to 31 in RuATD). This claim can be validated by splitting evaluation scores of the best binary RuATD models by length: on the texts longer than 23 words (about a quarter of all RuATD texts), top models can score over 0.95 accuracy.

Unsurprisingly all models that can outperform our BERT baseline used fine-tuned language models

(LMs) from the BERT family. Specific models that can achieve the best scores on the test set are mDe-BERTa (He et al., 2021), and Russian-language implementations of RoBERTa (Liu et al., 2019). Top models experiment with learning-rate scheduling as well as other training techniques (e.g., adversarial training with fast-gradient method (Dong et al., 2018), or child-tuning training (Xu et al., 2021)).

Using additional features (e.g., lexical richness, perplexity, number of characters, number of sentences, TF-IDF of POS tags, punctuation, tonality, reading ease) provided only limited benefit. While there are competitive solutions with such features (e.g., *mariananieva*, 4th-placed solution in the binary setup), none of the three best models in either task used any additional features.

Ensembling models proved to be beneficial, although competitive results could be achieved using single models. For example, *Posokhov Pavel*, the best model in the multi-class setup task, does not use the ensembling of any kind, nor does *orzhan*, the third-placed model in both tasks.

**Human Baseline** The overall accuracy of the human evaluation is 0.66, which scores below the BERT baseline. The low results are consistent with recent studies (Karpinska et al., 2021; Uchendu et al., 2021), which underpin the difficulty of the task for crowd-sourcing annotators. These works advise hiring experts trained to evaluate written texts or conduct multiple crowd-sourcing evaluation setups with extensive training phases. We leave the human evaluation experiments for future work.

## 6 Discussion

**On indistinguishable examples** The reasons for the errors of various systems on the RuATD corpus are of separate research interest. A short meaningful sentence of frequency n-grams may often occur in a web-corpus and be easily reproduced by a simple statistical LM. Thus, the very definition of a specific automatic text can be a challenging task for an attentive annotator and even for an engineer directly involved in developing TGMs. This can be illustrated, for example, by the case of Ilya Sutskever from the GPT-3 project, who tweeted spring of 2022, that *large neural networks may be "slightly conscious."*[20]. The methodological problem of obtaining some significant phrases or texts randomly using LMs, however, is raised much earlier than the onset of "indistinguishability by the engineers themselves": critical works on the Turing test (Turing and Haugeland, 1950) offer various variations of tests that level this problem. For example, (Bringsjord, 1996) explicitly note that a state machine that generates random sentences could be randomly considered meaningful by a judge in a good mood. In general, various methodological variations offer 1) interactive work with models/people, checking the maintenance of the context (Kugel, 1990) and even the consistency of the author's "cognitive profile" (Watt, 1996). These areas can be considered topics for future work for the following shared tasks.

**Ethical considerations** Setting the task of detecting non-human texts is timely due to the rapid development of LMs. The very issue of detecting non-human texts affects the fundamental right of the user to understand when they interact with a subjectless technological solution and when - with a person. Problems of this kind are actively discussed in reviews of recent years. In particular, (Bommasani et al., 2021) define the scope of problems as:

1. foundation model misuse, including both purposeful generated text misuse and the unconditional reliance on automatic text classification results that can be false negative;
2. development of legal grounds to mitigate generative model misuse and detection model misuse;
3. widespread deployment of automatic text detection systems: the presented models can lead to an "arms race" between malicious content generators and detectors.

Although the improvement of language modeling is undoubtedly a fundamental task of machine learning, we are of the position that a thorough study of models that classify automatic texts is necessary. As practice shows, the percentage of their errors in the Russian language is non-zero.

## 7 Related Work

Many research efforts are related to natural language generation (NLG) models. These works can be characterized into two broad categories - (i) training LMs on large-scale data and (ii) learning to dis-

---

[20]https://towardsdatascience.com

tinguish between machine-authored and human-written content. (Jawahar et al., 2020) provides a good survey on the automatic detection of machine-generated text for English.

Prior work has focused on training classifiers on samples from a model (Brown et al., 2020) and directly using a model distribution (Gehrmann et al., 2019). (Gehrmann et al., 2019) propose a visual and statistical tool named GLTR for the detection of generation artifacts across different sampling schemes. (Ippolito et al., 2020) compare human raters and automatic classifiers depending on the decoding strategy. They observe that classifiers can detect statistical artifacts of generated sequences while humans quickly notice semantic errors. Classifier accuracy ranges between 70% and 90% depending on the decoding strategy for short texts (64 tokens). (Dugan et al., 2020) propose a RoFT (Real or Fake Text) tool to detect the boundary between a human-written text passage and machine-generated sentences showing NLG models are capable of fooling humans by one or two sentences. A recent study of (Gallé et al., 2021) focuses on the unsupervised detection of machine-generated documents leveraging repeated higher-order n-grams. They show that specific well-formed phrases over-appear in machine-generated texts as compared to human ones. (McCoy et al., 2021) propose a suite of analyses called RAVEN for assessing the novelty of generated text, focusing on sequential structure (n-grams) and syntactic structure. Experiments show that random sampling result in generated text with a more significant number of novel n-grams.

Recent studies (Carlini et al., 2022; Lee et al., 2022) have raised a concern about model memorization due to data privacy leakage. (Carlini et al., 2022) confirm that memorization scales with model size and current LMs do accurately model the distribution of their training data. (Lee et al., 2022) investigate memorization and plagiarism when generating artificial texts. They observe that fine-tuned LMs demonstrate different patterns of plagiarism based on characteristics of auxiliary data. (Schuster et al., 2020) propose two benchmarks demonstrating the stylistic similarity between malicious and legitimate uses of LMs.

(Liyanage et al., 2022) propose a benchmark for detecting automatically generated research content that consists of a synthetic dataset and a partial text substitution dataset. The latter is created by replacing several sentences of abstracts with sentences generated by an NLG model. (Stiff and Johansson, 2021) adopt a wide variety of datasets of news articles, product reviews, forum posts, and tweets and investigated several classifiers to predict whether a text has been automatically generated. Their experiments show that classifiers perform reasonably accurately in the news domain, while the same task is more challenging for shorter social media posts.

## 8 Conclusion

We presented RuATD shared task, the first shared task on artificial text detection for the Russian language. As a result of the competition, 38 solutions have been obtained, solving the problem in two tasks modeled after the traditional concepts of the Turing test and authorship attribution for NLG methods.

The best solution of the shared task has achieved
- 83.0% accuracy in a binary task setup;
- 65.0% accuracy in a multi-class task setup.

The shared task dataset, codebase, human evaluation results, participant solutions, and other materials are now available online under Apache 2.0 license[21].

The competition problem can be further treated as a Turing test in a non-interactive setting. First of all, its direct methodological extensions are possible in such applied areas as:
- dialogue systems, and
- applications for editors and writers.

Another direction for future work is to conduct a critical study on the human evaluation guidelines on artificial text detection, which is still an open methodological question in the field (Karpinska et al., 2021). We welcome the communities of NLP developers, linguists, and engineers to contribute to further research in the area and next criteria formulations.

---

[21] https://github.com/dialogue-evaluation/RuATD

## Acknowledgments

## References

David Ifeoluwa Adelani, Haotian Mai, Fuming Fang, Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. 2020. Generating sentiment-preserving fake online reviews using neural language models and their human-and machine-based detection. // *International Conference on Advanced Information Networking and Applications*, P 1341–1354. Springer.

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

Anton Bakhtin, Sam Gross, Myle Ott, Yuntian Deng, Marc'Aurelio Ranzato, and Arthur Szlam. 2019. Real or fake? learning to discriminate machine from human generated text. *arXiv preprint arXiv:1906.03351*.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the opportunities and risks of foundation models.

S Bringsjord. 1996. The inverted turing test is provably redundant. *Psycoloquy*, 7(29).

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. // *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)*, P 7282–7296, Online, August. Association for Computational Linguistics.

Rosa María Coyotl-Morales, Luis Villaseñor-Pineda, Manuel Montes-y Gómez, and Paolo Rosso. 2006. Authorship attribution using word sequences. // *Iberoamerican Congress on Pattern Recognition*, P 844–853. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. // *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)*, P 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. 2018. Boosting adversarial attacks with momentum. // *Proceedings of the IEEE conference on computer vision and pattern recognition*, P 9185–9193.

Liam Dugan, Daphne Ippolito, Arun Kirubarajan, and Chris Callison-Burch. 2020. Roft: A tool for evaluating human detection of machine-generated text. // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, P 189–196.

Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. 2021. Tweepfake: About detecting deepfake tweets. *Plos one*, 16(5):e0251415.

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

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

Alena Fenogenova. 2021b. Text simplification with autoregressive models.

Matthias Gallé, Jos Rozen, Germán Kruszewski, and Hady Elsahar. 2021. Unsupervised and distributional detection of machine-generated text. *arXiv preprint arXiv:2111.02878*.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. 2019. Gltr: Statistical detection and visualization of generated text. // *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, P 111–116.

Ilya Gusev. 2020. Dataset for automatic summarization of russian news. // *Artificial Intelligence and Natural Language*, P 122–134, Cham. Springer International Publishing.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. // *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, P 4693–4703, Online, August. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

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

Vitaly Ivanin, Ekaterina Artemova, Tatiana Batura, Vladimir Ivanov, Veronika Sarkisyan, Elena Tutubalina, and Ivan Smurov. 2020. Rurebus-2020 shared task: Russian relation extraction for business. // *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog"]*, Moscow, Russia.

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

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

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A Conditional Transformer Language Model for Controllable Generation. *arXiv preprint arXiv:1909.05858*.

Pavle Kostenetskiy, Roman Chulkevich, and Viacheslav Kozyrev. 2021. HPC resources of the Higher School of Economics. // *Journal of Physics: Conference Series*, volume 1740, P 012050. IOP Publishing.

Peter Kugel. 1990. Is it time to replace turing's test?'. // *1990 Workshop Artificial Intelligence: Emerging Science or Dying Art Form. Sponsored by SUNY Binghamton's Program in Philosophy and Computer and Systems Sciences and AAAI.*

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. // *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, P 635–649, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Jooyoung Lee, Thai Le, Jinghui Chen, and Dongwon Lee. 2022. Do language models plagiarize? *arXiv preprint arXiv:2203.07618*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. // *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, P 110–119, San Diego, California, June. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Qi Liu, Matt J Kusner, and Phil Blunsom. 2020. A survey on contextual embeddings. *arXiv preprint arXiv:2003.07278*.

Vijini Liyanage, Davide Buscaldi, and Adeline Nazarenko. 2022. A benchmark corpus for the detection of automatically generated text in academic publications. *arXiv preprint arXiv:2202.02013*.

R Thomas McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. 2021. How much do language models copy from their training data? evaluating linguistic novelty in text generation using raven. *arXiv preprint arXiv:2111.09509*.

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

Michail Melnichenko and Natalia Tyshkevich. 2017. Prozhito from manuscript to corpus. *ISTORIYA*, 8(7 (61)).

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Andrey Sakhovskiy, Alexandra Izhevskaya, Alena Pestova, Elena Tutubalina, Valentin Malykh, Ivan Smurov, and Ekaterina Artemova. 2021. Rusimplesenteval-2021 shared task: evaluating sentence simplification for russian. // *Proceedings of the International Conference "Dialogue*, P 607–617.

Tal Schuster, Roei Schuster, Darsh J Shah, and Regina Barzilay. 2020. The limitations of stylometry for detecting machine-generated fake news. *Computational Linguistics*, 46(2):499–510.

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

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, P 8051–8067, Online, November. Association for Computational Linguistics.

Tatiana Shavrina and Olga Shapovalova. 2017. To the methodology of corpus construction for machine learning:«taiga» syntax tree corpus and parser. *Proceedings of the "Corpora*, P 78–84.

Harald Stiff and Fredrik Johansson. 2021. Detecting computer-generated disinformation. *International Journal of Data Science and Analytics*, P 1–21.

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

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

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

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

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

Stuart Watt. 1996. Naive psychology and the inverted turing test. *Psycoloquy*, 7(14):463–518.

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 S. Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from language models. *CoRR*, abs/2112.04359.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, P 38–45, Online, October. Association for Computational Linguistics.

Runxin Xu, Fuli Luo, Zhiyuan Zhang, Chuanqi Tan, Baobao Chang, Songfang Huang, and Fei Huang. 2021. Raise a child in large language model: Towards effective and generalizable fine-tuning. *arXiv preprint arXiv:2109.05687*.

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

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

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018. Generating Informative and Diverse Conversational Responses via Adversarial Information Maximization. *Advances in Neural Information Processing Systems*, 31:1810–1820.

# A    Annotation Protocols

**Задача. Человек или ИИ?**

Вам необходимо выбрать:

      1) текст написан человеком

      2) текст написан искусственным интеллектом (ИИ).

**Задание подробнее**

- Прочтите предложенный текст.
- Определите: текст написан человеком или ИИ?
- Если вы считаете, что текст написал человек, — выберите ответ "**ЧЕЛОВЕК**".
- Если вам кажется, что текст сгенерировал искусственный интеллект, — выберите ответ "**ИИ**".

**Примеры**

Текст: *А это моя семья моя семья моя семья.*

Определите: текст написан ИИ или человеком?

Нажмите "**ИИ**". Этот текст создан ИИ: текст будто "зациклен" на повторе *"моя семья"*.

Текст: *Софийский собор в Киеве, построенный в XI веке, является одним из самых значительных архитектурных сооружений данного периода. Первоначально Софийский собор представлял собой пятинефный крестово-купольный храм с 13 главами. С трёх сторон он был окружён двухъярусной галереей, а снаружи — ещё более широкой одноярусной.*

Определите: текст написан ИИ или человеком?

Нажмите "**ЧЕЛОВЕК**". Это написал человек. Связность текста и исторические факты не нарушены.

**Обратите внимание!**

Вам будут встречаться тексты разной длины и разных видов: это могут быть новости, официальные документы, тексты из социальных сетей или художественные. По следующим особенностям можно определить, что текст написан ИИ:

- Несогласованность фактов и/или потеря связности текста
- Нарушение логической связи или знаний о мире
- Бессмысленные повторы или внезапное по смыслу завершение текста

Следующие признаки могут ввести в заблуждение, так как иногда встречаются и у людей и у машин:

- Грамматические/орфографические ошибки. Иногда даже люди совершают ошибки и опечатки, это нормально.
- Стиль. Современный ИИ может имитировать стиль текста и использовать разнообразную лексику. Если текст написан правильно, это необязательно человек!

Спасибо!

Figure 1: An example of the annotation instruction for the human evaluation.