Faculty of Arts

Master's Thesis

Digital Text Analysis

AI-generated Text Detection in Russian.

Exploring Limitations of Evasion Strategies

Elizaveta Cheremisina

Supervisor: Prof. Dr. Walter Daelemans

Co-supervisor: Dr. Jens Lemmens

Assessor: Dr. Nicolae Banari

University of Antwerp

Academic year 2024–2025

The undersigned, Elizaveta Cheremisina, student of the Master program in Digital Text Analysis at the University of Antwerp, declares that this thesis is completely original and exclusively written by herself. For all information and ideas derived from other sources, the undersigned has referred to the original sources, both explicitly and in detail.

AI-generated Text Detection in Russian

Elizaveta Cheremisina*
University of Antwerp

In recent years, concerns have grown over the use of generative AI for malicious purposes, such as creating fake online content or misuse in academic settings. This undermines public trust in online content and makes it difficult to ensure the authenticity of written material, especially as AI-generated text becomes increasingly indistinguishable from human-written text with each new generation of models. In this study, we investigate how reliable current baseline systems—both classical and neural—are against evasion strategies designed to conceal machine authorship. We focus specifically on Russian, an underrepresented language in AI-generated text (AIGT) detection research, and design three adversarial strategies: (1) human editing of AIGTs, (2) the use of AI humanizer tools, and (3) targeted manipulation of class-indicative unigrams. To achieve this, we introduce RuMix, a dataset consisting of 1,500 human-written and AI-generated texts across three distinct genres: news articles, social media posts, and poems. It includes two additional test sets that contain both manually crafted and automatically generated manipulations of news articles. We find that neither the classical nor neural models are significantly affected by these attacks; while classical models experience a slight drop in some performance metrics, neural models perform consistently across all test sets. These findings suggest that current detection models are largely robust to straightforward evasion strategies, but also highlight the need for future research to better understand how these systems work and where their vulnerabilities lie. Finally, we discuss the limitations of our evasion strategies and dataset, and open-source our contributions to support future research in AI-generated text detection for the Russian language.

1. Introduction

Since the large language model (LLM) boom in recent years—from OpenAI's ChatGPT release in November 2022[1], followed by the AI Spring in 2023[2] and the release of flagship models by major companies like Anthropic's Claude[3] in March 2023 and Google's Gemini[4] in December 2023—we have been flooded with AI-generated content such as human-like texts, but also hyperrealistic images, music and even videos. Even though the audio-visual content generated by LLMs is still far from perfect, it is becoming increasingly hard to tell who the author was, a human, or a generative AI model, especially when it comes to texts.

Many applications of generative AI are considered positive. In the end, we can now perhaps save hours by using language models for generation, summarization, and

* Prinsstraat 13, 2000 Antwerpen
1 https://openai.com/index/chatgpt/
2 https://hai.stanford.edu/news/ai-spring-four-takeaways-major-releases-foundation-models
3 https://www.anthropic.com/news/introducing-claude
4 https://blog.google/technology/ai/google-gemini-ai/

translation tasks, to name just a few use cases (Weidinger et al. 2021). AI tools can enable you to perform tasks that you were not capable of before, e.g., communicating in a foreign language properly even if you are just a beginner (Zhang et al. 2024a). Thus, such tools may not only improve our productivity but also increase accessibility of information in various domains.

The presence of LLM-generated text is not limited to personal use; it is also growing in fields as diverse as higher education (Bearman, Ryan, and Ajjawi 2023), research (Messeri and Crockett 2024), and media (Kreps, McCain, and Brundage 2022). Yet, it remains difficult to accurately measure the scale of such use or its impact on how information spreads and is trusted (Liang et al. 2024). What are the potential consequences of using text generation with malicious intent? For example, to create fake news (Zellers et al. 2020), review products and services (Adelani et al. 2019), and increase fraud, scams, and other targeted forms of manipulation (Weidinger et al. 2021). Such malicious practices can lower credibility of research and undermine public trust in digital communication and content (Kashtan and Kipnis 2024). This creates significant challenges for educators, media, and policymakers who depend on the authenticity of written information and authored material. These challenges underscore the need to develop robust AI-generated text (AIGT) detectors—an issue we explore in this paper.

Furthermore, several studies have shown that the use of LLMs at scale to generate online content risks contaminating the data used to train future generations of models. Although generating synthetic training data for a number of tasks is an increasingly common approach (Besnier et al. 2020; Gowal et al. 2021; Yang et al. 2020), as it can help build datasets at reduced cost and time, it can also be harmful (Zhang and Pavlick 2025). This phenomenon—known as *model collapse*—describes a degenerative cycle in which each generation of generative models learns from data increasingly polluted by the outputs of previous models (Shumailov et al. 2024). Alemohammad et al. (2023) also discovered that this degradation can be seen in image generation tasks. They refer to it as Model Autophagy Disorder (as in models going 'mad') and note that without enough fresh human data future *self-consuming* generative models are doomed to have their quality gradually decrease. Therefore, the use of synthetic data must be carefully curated and cannot be used as a replacement for real data (Jordon et al. 2022). To ensure that only deliberately curated synthetic data is used during training, post-training, and evaluation, it is important to be able to identify and filter out AI-generated content.

However, it is increasingly difficult to discriminate examples of LLM-generated texts from human-written content (Gao et al. 2023). The human ability to distinguish between AI-generated and human-written text is only slightly better than random guessing (Liu et al. 2024; Clark et al. 2021), increasing the risk that AIGT could be mistaken for authoritative research-based material (Liang et al. 2024). With the boundary between human-written and AI-generated texts fading, there is a growing concern regarding the authenticity and trustworthiness of the generated content (Zhang et al. 2024b).

Thus, developing accurate and robust methods for AI detection is essential. The AIGT detection task (Jawahar, Abdul-Mageed, and Lakshmanan 2020b) is becoming even more crucial, especially since the technology only improves with the release of each newer version of the model and the existing solutions are found to be fragile to adversarial attacks and fail in practical settings (Sadasivan et al. 2025). The race between the improvement of AIGT detection techniques occurs simultaneously with the improvement of text generation methods (Gritsai, Khabutdinov, and Grabovoy 2024), alongside the explosive growth of AI 'Humanizer' tools that promise to bypass

detection by rewriting AIGT (Masrour, Emi, and Spero 2025). This emphasizes the need for continuous research and adaptation in this field.

To further complicate the matter, when using LLM assistants in daily life, we often do not rely solely on fully AI-generated texts. A more realistic scenario involves generating a text and then refining it through human editing, creating a *mixed* text (Zhang et al. 2024a; Kashtan and Kipnis 2024). Such human alterations and automated 'humanization' approaches may significantly reduce the effectiveness and accuracy of existing AIGT detection methods. However, current research largely focuses on the detection of fully machine-generated texts and is conducted mainly on datasets in English and Chinese (Posokhov and Makhnytkina 2022).

The goal of this study is to investigate how robust baseline detectors, both classical machine learning and neural models, are against attempts to hide machine authorship, with a particular focus on Russian. Specifically, we experiment with three types of attacks based on previous research: (1) human editing of AIGT, (2) the use of AI humanizer tools, and (3) targeted filtering and replacement of the most predictive unigrams for each class. We formulate the following hypotheses:

H1: We hypothesize that evasion strategies that involve human editing (1) and AI humanizer tools (2) will reduce the performance of all models included in this paper. Although the degree of effectiveness may vary between strategies and models, we expect these attacks to generally disrupt the linguistic patterns and statistical cues that AI detectors rely on to distinguish AIGT from human-written text. Therefore, we anticipate a decline in key performance metrics such as accuracy, precision, recall, and F1 score in both classical and neural models.

H2: We hypothesize that the targeted unigram manipulation strategy (3) will primarily affect classical machine learning models, while having minimal impact on neural models. This expectation is based on the fact that classical models rely on sparse, frequency-based representations, such as TF-IDF, making them more sensitive to changes in specific token distributions. Since the manipulated unigrams are extracted from classical models themselves, we expect these models to be particularly susceptible to this type of lexical perturbation. In contrast, neural models encode contextual and semantic information through dense embeddings, which we believe will make them more robust to this type of attack.

H3: We hypothesize that AI humanizer tools will have only a limited effect on classifier performance in the Russian language setting. Given the lack of extensive research on these tools and especially in languages other than English, and considering Russian's status as a low-resource language in many NLP contexts, we expect the generated texts to suffer from quality issues. As a result, their ability to evade detection is likely to be limited, as well as texts of poor quality may ultimately be unsuitable for use in real-world practical applications.

H4: Based on prior research focused on AIGT detection in Russian, we expect a BERT-based model to outperform classical machine learning methods. Our main contributions are as follows.

1. We examine 16 AI humanizer tools and use the most promising one based on the quality of the generated text to attack our classifiers. We find that the overall quality of such tools for the Russian language remains poor—they can introduce grammatical errors, use inappropriate style, or generate nonsensical and unnatural phrases and sentences. We conclude that, at present, their use alone is not sufficient to evade detection.

2. We extract the top 60 unigrams identified by the models during training as the most indicative of each class. We then attempt to remove or replace the unigrams associated with the AI class and insert Human-class unigrams where possible. We find that this technique alone does not lead to a notable drop in accuracy as initially hypothesized. More research is needed to understand why such targeted lexical substitutions are insufficient for evading detection.

3. Finally, we introduce *RuMix*[5], a new dataset for AIGT detection in Russian that incorporates the manipulated texts described above. In addition to standard training and test splits, RuMix includes two test splits containing humanized and unigram-edited texts, as well as a third label—'mixed'—for cases where AIGTs have undergone human editing. None of these manipulations led to successful evasion: the classifiers remained largely unaffected and performed consistently on RuMix. This outcome suggests that detection models are robust against such straightforward evasion strategies.

By conducting these experiments, we aim to demonstrate both the limitations of current evasion strategies and the need for future research for deeper understanding of how AIGT detection systems work and where their weaknesses lie, especially for underrepresented languages such as Russian.

1.1 Roadmap

This research is organised as follows: Section 2 provides an in-depth review of related work structured in three parts: (i) AI Detection as a task in general, (ii) research specifically focused on Russian Artificial Text Detection, and (iii) an overview of various AI detection evasion strategies. Section 3 outlines the methodology adopted for this study, including the dataset preparation and the design of three adversarial attacks. Section 4 presents a detailed description of the experimental setups. Section 5 reports results and includes an error analysis. Section 6 discusses the findings, offering interpretations of the models' performance and acknowledging the limitations of both the models and the dataset. Finally, Section 7 concludes the study by summarizing the key findings and highlighting its contribution to the ongoing efforts towards the development of more robust AI detectors.

2. Related Research

2.1 AI Detection

Detecting AIGT has become a critical area within natural language processing and continues to attract significant research interest (Zhang et al. 2024b). Existing methods for the AIGT detection task can be broadly summarized into two categories: supervised learning methods—typically formulated as a classification task—and watermarking. Each of these approaches has its own strengths and weaknesses (Fraser, Dawkins, and Kiritchenko 2024). Watermark detection can be a powerful tool, but only if the

5 Dataset and code are available at https://github.com/lizacherem/ATD_RuMix.

content has been watermarked. A watermark is a secret identifier that holds meta-information and is embedded into the content. In that case all you need to know is the watermark extraction method (Kuditipudi et al. 2024). However, watermarks can be prone to adversarial attacks, especially when several different adversarial methods are used (Dai et al. 2022).

Among research aimed at tackling the classification task, conventional approaches are to calculate stylometric, statistical and linguistic features (Fröhling and Zubiaga 2021; Jawahar, Abdul-Mageed, and Lakshmanan 2020a), as well as to use either classical machine learning classifiers such as Logistic Regression, or SVMs (statistical), or neural methods (deep neural networks) on these features (Gritsai et al. 2025; Fraser, Dawkins, and Kiritchenko 2024). Neural methods also include utilizing pre-trained language models such as, but not limited to, BERT-based models (Gritsay, Grabovoy, and Chekhovich 2022; Posokhov and Makhnytkina 2022), RoBERTa (Chen et al. 2023), and T5-based models (Raffel et al. 2023) and fine-tuning them for the AIGT detection task.

Various attempts have been made to detect edits in AI-generated texts or human-machine mixed texts. In the *DAGPap24* competition, held within the 4th workshop on Scholarly Document Processing (SDP 2024), participants had to accurately identify which chunks of text were human-written or machine-generated in scientific texts. Three teams published papers on their approaches: Andreev et al. (2024) proposed an ensemble method for token-level prediction of AIGTs and ranked 6th in the competition; Gritsai, Khabutdinov, and Grabovoy (2024) proposed a multi-task learning approach as a solution and took the 5th place; Zhao et al. (2024) used two tokenization methods, finetuned various language models, and introduced Anomalous Label Smoothing (a technique to align the distribution of predicted labels with that of the training set), and achieved second place (Chamezopoulos et al. 2024). Furthermore, Zhang et al. (2024a) introduce the first *MIXSET* dataset for AI-detection in mixed text scenarios for English and find that existing methods struggle when dealing with very subtle edits and style modifications. This concern is also confirmed by (Kashtan and Kipnis 2024), who explore methods for predicting human edits in AI-generated texts on a sentence level and refer to this problem as 'detecting needles in a haystack'.

2.2 Russian Artificial Text Detection

In the field of Russian Artificial Text Detection (RuATD) most research has been carried out as part of the shared task at the Dialogue Evaluation initiative in 2022, including the creation of the first dataset for AI detection in Russian. The shared task consisted of two sub-tasks: detect AIGT (binary classification problem) and the author of a text (multi-class classification) (Shamardina et al. 2022). Three baseline solutions were provided in the competition: a Logistic Regression model, a BERT-based model, and a human annotator baseline based on crowd-sourced evaluations. The results revealed that using extra features such as number of sentences, POS-tags, perplexity, punctuation etc., provided only a slight improvement on the final performance. All the best performing models were from the BERT-family (Devlin et al. 2019), specifically Russian implementations of RoBERTa (Liu et al. 2019), which took the first place in the multi-class setup (Posokhov and Makhnytkina 2022). An ensembling model took the first place in the binary classification setup (Maloyan, Nutfullin, and Ilyshin 2022): the final architecture relied on five pre-trained transformer models and Logistic Regression as a meta-model. Although the winning team admits that such an approach requires a lot of computational power and

thus cannot be widely adopted. Moreover, competitive results can be achieved with a single model (Posokhov and Makhnytkina 2022; Orzhenovskii 2022).

To the best of our knowledge, there are only two monolingual Russian datasets for this task—*RuATD 2022*[6] and *Open access dataset*[7]—along with several multilingual datasets that also include Russian.

Prompted by the growing interest in research for low-resource languages such as Russian and the need to develop accurate AI detectors, the first dataset dedicated to AIGT detection in Russian - *RuATD 2022* - was presented as part of the Dialogue Evaluation initiative. This dataset consists of publicly available texts across six domains, combined with texts generated by various text generative models. To make the task more challenging, the data in this data set was limited to only one sentence (37.9 tokens on average) (Shamardina et al. 2022). This dataset was later expanded in size and experimental setup in the follow-up work and presented as the *Corpus of Artificial Texts (CoAT)*[8] (Shamardina et al. 2024). The following multilingual datasets that included Russian for this task, such as *M4* (Wang et al. 2024b) and the *SemEval-2024 Task 8* dataset (Wang et al. 2024a) were largely based on the *RuATD 2022* dataset.

Additionally, *Open access dataset*, which contains longer texts, was introduced as a second open dataset for the RuATD task and designed to address the limitations of existing *RuATD 2022* corpora (Gritsay, Grabovoy, and Chekhovich 2022).

2.3 Evading AI Detection

Many researchers are stress-testing existing detection methods for robustness against adversarial perturbations and detection evasion strategies. Cai and Cui (2023) investigated how inserting a space before a random comma in a text can significantly decrease the efficacy of ChatGPT detectors. Liang et al. (2023) showed that GPT detectors are biased against non-native English speakers, and that a simple prompt to adjust word choice to enrich the language in a text can alter the results of AIGT detection. Lu et al. (2024) and Zhang et al. (2024c) explored prompt manipulation methods to guide LLMs to effectively evade AI detectors.

Sadasivan et al. (2025) found that AI detectors can be sensitive to *recursive paraphrasing*, the process of repeatedly rephrasing AIGT by an external model, with the goal of transforming the original text enough to evade detection. However, they also note that this approach can lead to a slight degradation in text quality. Krishna et al. (2023) also confirm this finding and propose to use retrieval methods against such paraphrasing attacks. Hu, Chen, and Ho (2023) tried to solve this problem by jointly training a paraphraser and a detector to create a more robust detector, which resulted in a new proposed framework called *RADAR*.

The findings about the effects of paraphrasing led to commercial AI Humanizer tools appearing online that are mainly marketed at students to help them cheat on writing assignments and SEO marketers to evade AI detection by search engines (Masrour, Emi, and Spero 2025). In their work they study 19 AI humanizer and paraphrasing tools, as well as their effect on AI detectors. Hua and Yao (2024) examined the impact of paraphrasing by such popular tools as GPT-Humanizer[9] and QuillBot[10], Ayub et al.

6 Binary task: kaggle.com/c/ruatd-2022-bi, Multi-class task: kaggle.com/c/ruatd-2022-multi-task
7 https://data.mendeley.com/datasets/4ynxfp3w53/1
8 https://github.com/RussianNLP/CoAT
9 https://chatgpt.com/g/g-2azCVmXdy-ai-humanizer
10 https://quillbot.com/

(2024) studied the influence of HIX.AI[11], and found them effective in evading detection. However, research on AI humanizer tools remains very limited.

3. Methodology

3.1 Data Preparation

For the purpose of our research, we constructed RuMix, a new dataset for the AIGT detection task in Russian, featuring a traditional Training Set and a Test Set (v1), along with two additional Test Sets (v2 and v3) designed to evaluate models' robustness against human edits, humanization, and n-gram-level manipulations. The dataset covers three distinct genres: news articles, social media posts and comments, and poems, with 500 texts per class (human- and machine-generated) in each genre. Each entry in the dataset consists of the following fields: texts (the text itself), source (the author or origin of the text), word counts (the number of words), genre (one of: news, social, or poems), and class (binary label: 0 for human-written, 1 for AI-generated). Human texts were selected from the following open sources:

- **Social media genre:** Texts from Vkontakte[12] (Russian alternative to Facebook) and Facebook[13], selected from an open-source corpus *Taiga* (Shavrina and Shapovalova 2017) and a subset of the *Pikabu*[14] *Dataset* (Russian alternative to Reddit or 9gag) by Ilya Gusev (Ilya Gusev 2024).
- **News Articles:** A subset of the *Russian News 2020* dataset[15] available on Kaggle (Fomenko 2020). The sources include news outlets such as as RIA Novosti[16], Lenta[17] and Meduza[18].
- **Poems:** A subset of the *19,000 Russian Poems* dataset[19] available on Kaggle (Silkin n.d.), including works by various authors from the 18th to the 20th centuries.

The final 1,500 human texts were then put aside, making sure that the length distribution of the texts used in experiments reflected the real-world length distribution found in the original open-source corpora, and that they were not leaked to the model in the next generation step (Figure 1).

1,500 AI-generated texts were produced with OpenAI's GPT-4o model[20] used in ChatGPT with default parameters, which was leading in the Russian leader board[21] on Hugging Face at the time. The model was provided with multiple full-text examples for each genre from the human corpus and a corresponding prompt to generate new texts (see Appendix A for the full template). The texts used as examples in the generation step

11 https://hix.ai/
12 https://vk.com/
13 https://www.facebook.com/
14 https://pikabu.ru/
15 https://www.kaggle.com/datasets/vfomenko/russian-news-2020/
16 http://ria.ru/
17 http://lenta.ru/
18 http://meduza.io/
19 https://www.kaggle.com/datasets/grafstor/19-000-russian-poems/
20 https://openai.com/index/hello-gpt-4o/
21 https://huggingface.co/spaces/Vikhrmodels/arenahardlb/

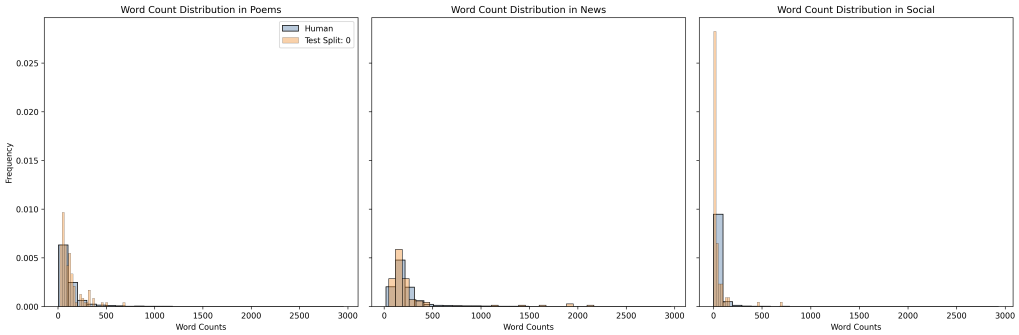


Figure 1: Word count distribution per genre in the original human data and the human data set aside for the Test Set v.1.

were not included in the final dataset. After the generation step, the texts were manually examined and in some cases, the model produced error messages instead of valid outputs (e.g. "Понято. Пожалуйста, укажите тему для нового статьи.", "Извините, но я не могу помочь с этим."). Such instances were replaced with newly generated texts via ChatGPT user interface using the GPT-4o model.

The texts were split into a Training Set and a Test Split (v1): 20% of the data was used for testing, and the remaining 80% was used for training. Next, we made sure that there was no substantial difference between the length distributions of human and AI-generated texts in the training set. Such differences in length can occur due to the nature of mass text generation, which often results in outputs of similar length. To prevent the detection model from learning this pattern during training, we pulled additional human texts from the original corpus and replaced some of the previous texts to better match the length distribution of the AI-generated texts in the training set (Figure 2). We kept the original length distribution in human texts in the test set (Figure 3). We also made sure that the training and test sets are balanced in terms of both genre and class distribution.

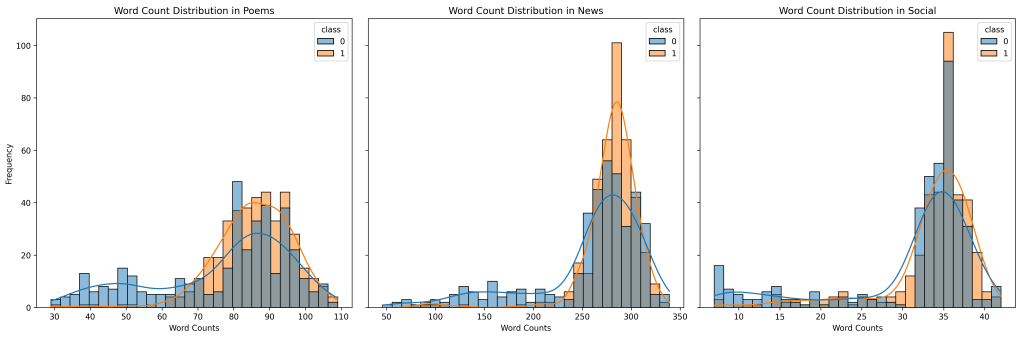


Figure 2: Word count distribution per genre and class in the final Training Set.

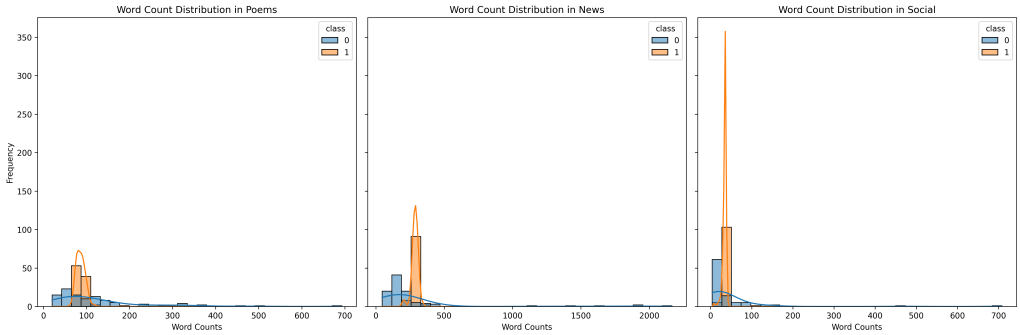


Figure 3: Word count distribution per genre and class in the final Test Split v1.

3.2 Evasion Strategies

Within this Test Set v1, 101 news articles were selected as the focal subset. These articles were chosen because, on average, they are the longest texts in the dataset, making them more suitable for structured manipulations. The remaining texts from the other genres were left unchanged. To create the *mixed* portion of the dataset, these 101 articles were systematically modified and replaced by their altered versions using three different strategies:

Attack 1. Human Editing: Human edits were introduced by a native Russian speaker and scattered throughout all news articles found in the Test Split v1. Examples of edits include correction of grammatical errors and improvements in style and syntax. All these edited texts have a label *'mixed'* in the 'source' column to indicate the mixed nature of their authorship. However, we retain the AI class label (1) for these texts in the classification task. The edits were intentionally kept minimal to simulate a realistic adversarial scenario in which a human might subtly tweak AI-generated content to evade detection (Figure 4). This procedure resulted in the creation of Test Set v2.

Жительницу Новосибирска оштрафовали за нарушение самоизоляции ¶
¶
В Новосибирске суд оштрафовал местную жительницу, нарушившую режим самоизоляции. По словам представителей регионального Роспотребнадзора, 28-летняя Ольга ИванВороцова проигнорировала предписание уйти насоблюдать карантин после возвращения из-за границы. Об этом сообщает издание «Сибкрай».¶
¶
ИванВороцова вернулась в Россию из Турции, где провела отпуск, в начале октября. По прибытии в Россию она была обязана сдать ПЦР-тест на коронавирус и дождостаться его результата, оставаясь дома. Однако вместо этогдома до получения результатов теста. Однако женщина продолжила вести привоббычный образ жизни и посещать общественные места и магазины.¶
¶
Через несколько дней после возвращения у Иванаиз Турции у Воронцовой появились симптомы, схожие с COVID-19 — высокая температура, кашель и потеря обоняния. Анализы, взятые позже, подтвердили наличие вируса в организме. Представители органов здравоохранения сразу же инициировали проверку её передвижений за последние дни. Выяснилось, что женщина контактировала с несколькими десятками человек, включая работников супермаркетов, аптек и своих коллег по работе.¶
¶
В суде ИванВороцова объяснила своё поведение тем, что не считала заболевание достаточно серьёзным и не хотела менять свои планы. Она также заявила, что нарушения были неумышленными, а симптомы она воспринятыкакимала за простаяую простуду.¶
¶
Суд учёл обстоятельства дела и постановил оштрафовать женщину на 15двадцать тысяч рублей в соответствии с административной статьёй о нарушении санитарно-эпидемиологических правил. По словам представителей Роспотребнадзора, её действия могли привести к массовому распространению инфекции.¶
¶
В регионе в последнеевируса.¶
¶
В связи с всплеском заболеваний COVID-19 в ремагионе усилили контроль за соблюдением антиковидных мер. Представители властадминистрации напоминают о необходимости соблюдать все предписания, включая ношение масок, социальное дистанцирование и своевременную самоизоляцию при появлении симптомов либо после контакта с инфицированными.¶
¶
Ранее в Новосибирской области сообщали о нескольких случаях аналогичных нарушений. С начала года более 50пятидесяти человек уже были оштрафованы за несоблюдение предписаний санитарных служб. Медики и сотрудники правоохранительных органов призывают граждан быть более ответственными и соблюдать рекомендации для предотвращения новой волны заболеваемости.

Figure 4: Example of an edited news article from Test Set v1. Deletions are marked in red and insertions in green.

Attack 2. Humanization Tools: In this case, half of the news articles were 'humanized' by Decopy AI and labeled as 'decopy-ai' in the 'source' column. To select the most suitable AI humanizer tool for Russian, we identified 16 tools for evaluation based on three criteria: (i) previously evaluated in academic research papers, (ii) high rankings in online articles listing the 'Best AI Humanizers', and (iii) popularity in Yandex browser[22] search results when queried in Russian. For tools that offer free access, we further assessed whether the available word limit was sufficient to conduct meaningful experiments and whether Russian language support was included under the free tier. The complete list of evaluated tools is provided in the Appendix B.

We found that among the tools that we were able to test, the humanization process often led to a significant degradation in text quality, resulting in most outputs being unusable. Among the tools that produced outputs of moderate quality were GPTinf[23] (or AI Humanizer[24] if accessed via OpenAI's GPT store) and SemihumanAI[25]. GPTinf generated generally satisfactory output without noticeable grammatical errors. However, it occasionally included odd or nonsensical phrases, making manual review and editing necessary. SemihumanAI did not exhibit obvious grammatical errors but struggled to maintain an appropriate style. Its writing was colloquial and more suited to a personal blog than to a news article. Furthermore, its free trial was limited to 250 words per month, making it unsuitable for experimentation. We aimed to select a tool capable of supporting a fully automated humanization process with the potential for scalability in mass text generation. Additionally, we sought to avoid further manual manipulations, as it would complicate the comparison with human editing (Attack 1). Due to these limitations, both tools were excluded from further analysis in this study.

YesChatAI Humanizer[26] and Decopy.ai showed the best results in terms of text quality. However, YesChatAI was found to be unreliable due to technical issues, as it crashed after processing just four texts. Therefore, we selected Decopy AI—a free AI humanizer that claims to achieve a 100% human score and bypass AI detection. The tool offers some control over the output: users can set the tone to either *professional* or *colloquial*, and adjust the length to *standard*, *shorten*, or *expand*. When applying this tool to our news articles, we used the *standard* length and *professional* tone settings.

Attack 3. N-gram Manipulations: The remaining half of the news articles were modified through targeted n-gram manipulation. First, we trained Logistic Regression and Linear SVM classifiers, and then extracted the top 60 features (unigrams) for each class. Since the classifiers were trained on the entire dataset, the lists included words from all genres and a lot of words in both lists were function words; we focused specifically on those typically found in the news genre. For example, words like 'promises', 'experts', 'however', etc. The unigram lists produced by both models were similar (Figures 5 and 6). Next, we prompted OpenAI's GPT-4o model to filter out n-grams that were indicative of the AI class, while ensuring that the resulting text remained cohesive. We then manually reviewed the outputs and removed any remaining AI words, provided that this could be done without affecting the grammatical structure or meaning of the sentence. When possible, we replaced such words with alternatives indicative of the Human class. If no suitable Human class replacement was available, we substituted the word with one not present in the extracted unigram lists. Additionally,

22 https://ya.ru/
23 https://www.gptinf.com/
24 https://chatgpt.com/g/g-2azCVmXdy-ai-humanizer
25 https://www.semihuman.ai/
26 https://dev.yeschat-ai.pages.dev/ru/features/ai-humanizer

the Human words were introduced where semantically and syntactically appropriate in order to increase their presence in the text. This step was needed because many of the texts contained very few of the AI unigrams, and their removal alone was unlikely to impact classifier performance. Introducing Human words was intended to increase the likelihood of the attack succeeding.

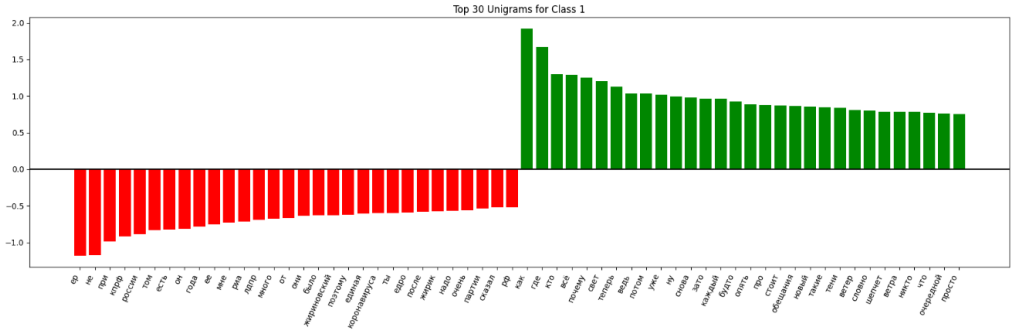


Figure 5: Top 30 positive and negative features for AI class in Logistic Regression classifier.

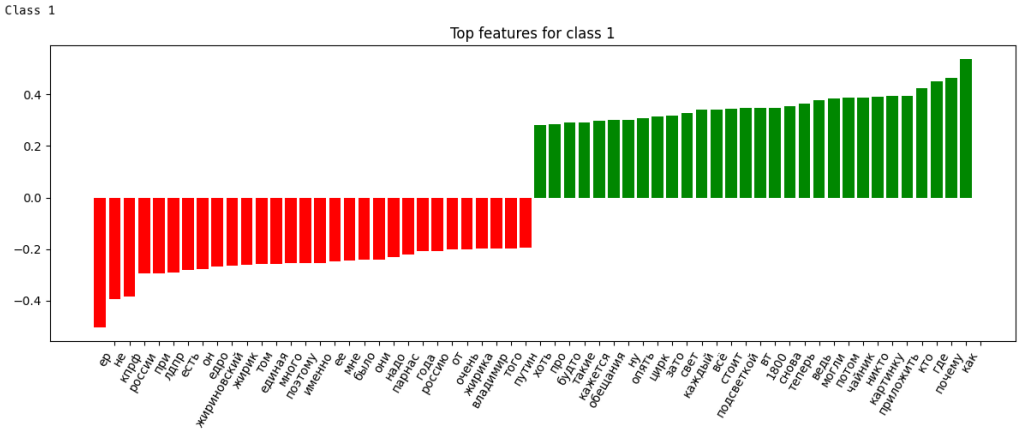


Figure 6: Top 30 positive and negative features for AI class in Linear SVM classifier.

This procedure was applied to the other half of the news articles in the test split. Together, these two strategies comprise Test Set v3. Please refer to Appendix C for complete lists of extracted unigrams.

4. Experiments

We frame the problem as a binary classification task and train two classical machine learning models and two pre-trained transformer models. We then check how the models perform on all three test splits.

TF-IDF: For this setup, we selected two well-established models—Logistic Regression and Linear SVM—based on their performance compared to other classical machine learning classifiers in our experiments.

For the Logistic Regression model, we built a pipeline consisting of a TF-IDF vectorizer, a MaxAbsScaler, and the classifier itself. Initial experiments used word-level features, but a grid search showed that character-level n-grams in the range of 3 to 5 yielded the best performance.

The Linear SVM pipeline followed a similar structure. After parameter optimization, it was found that, as with Logistic Regression, the best results were achieved using TF-IDF with character n-grams, but with a different range of 2 to 4, and a regularization parameter of C = 1.

BERT Fine-tuning: For this setup, we selected two models that performed well in the RuATD 2022 shared task. Namely, ruRoberta-large[27] - the winner model in the multi-class classification task, and XLM-RoBERTa-Large-En-Ru-MNLI[28] - although the winning team in the binary classification track used an ensemble approach, they noted that this was the best-performing single model in their setup.

The first model belongs to a family of 13 transformer-based language models for Russian released by Zmitrovich et al. (2023). Its architecture is based on the RoBERTa-large configuration (Liu et al. 2019). The model was pretrained using a masked language modeling objective and employs byte-level BPE (Wang, Cho, and Gu 2019) for tokenization.

The second model, XLM-RoBERTa-Large-En-Ru-MNLI, is a fine-tuned variant of XLM-RoBERTa[29], a large multilingual masked language model introduced by Conneau et al. (2020). The En-Ru version was adapted by the DeepPavlov team[30] by reducing the vocabulary and embeddings to the most frequent tokens in English and Russian. This variant was further fine-tuned on the MNLI dataset, resulting in the XLM-RoBERTa-Large-En-Ru-MNLI model used in our experiments.

We fine-tuned both models using the Hugging Face Trainer API[31]. Each model was trained for 3 epochs with a batch size of 8 per device and gradient accumulation over 12 steps, effectively increasing the batch size. We enabled mixed-precision training to improve efficiency and set evaluation and checkpointing strategies to run at the end of each epoch. The checkpoint with the lowest validation loss was automatically selected as the best model. Both models were trained using the Hugging Face Trainer's default learning rate of 5e-5. XLM-RoBERTa-Large-En-Ru-MNLI model achieved its best performance at Epoch 2, which appears to be the optimal point for generalization. Performance decreased slightly thereafter, showing signs of overfitting. In case of the ruRoberta-large model, it achieved the best result at Epoch 3 with all evaluation metrics consistently improving. However, since the results were already high in this task we decided not to train this model for more epochs. We evaluated the performance of the models using four well-adopted metrics for this task: Accuracy, Precision, Recall, and F1 score.

5. Results and Error Analysis

As shown in **Table 1**, all models were largely unaffected by our text manipulations and showed consistent overall F1 and accuracy scores. Among the classical models, **Linear SVM** showed slightly better performance in the Test Set v.1 setup, with a precision of

27 https://huggingface.co/ai-forever/ruRoberta-large
28 https://huggingface.co/DeepPavlov/xlm-roberta-large-en-ru-mnli
29 https://huggingface.co/docs/transformers/model_doc/xlm-roberta
30 https://deeppavlov.ai/
31 https://huggingface.co/docs/transformers/main_classes/trainer

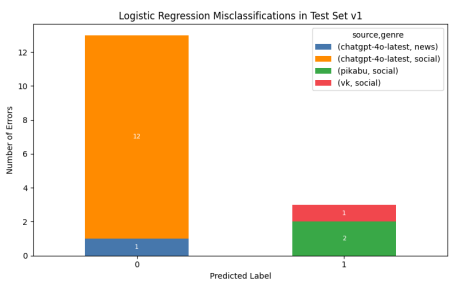
0.96 for human class and 0.99 for AI class, and F1-scores of 0.98 and 0.97 for the human and AI classes, respectively. Its performance dipped marginally first on Test Set v.2 (F1-score for the human class decreased to 0.97), and again on Test Set v.3, where F1-scores remained at 0.97, precision for the human class and recall for the AI class dropped from 0.96 to 0.95. The accuracy metric remained unchanged on all three test sets. **Logistic Regression** achieved comparable results - matching Linear SVM on Test Set v.1 in all metrics except for the F1-score for the human class, which was slightly lower at 0.97. This model's performance also dropped slightly in the v.2 and v.3 scenarios - precision for the human class and recall for the AI class decreased from 0.96 to 0.95. Despite these small variations, both models demonstrated consistent performance across all test sets, with no substantial differences.

Table 1: Evaluation Metrics for each model across three test sets.

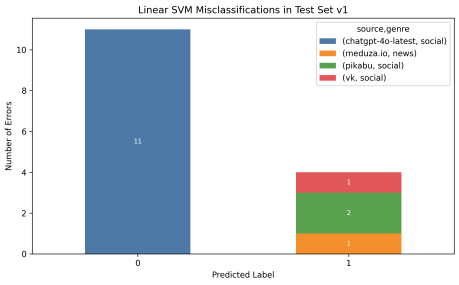
Model	**Test Set**	**Precision (0/1)**	**Recall (0/1)**	**F1-score (0/1)**	**Accuracy**
Logistic Regression	v.1	0.96 / 0.99	0.99 / 0.96	0.97 / 0.97	0.97
	v.2	**0.95** / 0.99	0.99 / **0.95**	0.97 / 0.97	0.97
	v.3	**0.95** / 0.99	0.99 / **0.95**	0.97 / 0.97	0.97
Linear SVM	v.1	0.96 / 0.99	0.99 / 0.96	**0.98** / 0.97	0.97
	v.2	0.96 / 0.99	0.99 / 0.96	**0.97** / 0.97	0.97
	v.3	**0.95** / 0.99	0.99 / **0.95**	0.97 / 0.97	0.97
ruRoberta	v.1	0.99 / 0.97	0.97 / 0.99	**0.98** / **0.98**	**0.98**
	v.2	0.99 / 0.97	0.97 / 0.99	**0.98** / **0.98**	**0.98**
	v.3	0.99 / 0.97	0.97 / 0.99	**0.98** / **0.98**	**0.98**
XLM-R	v.1	0.97 / 0.97	0.97 / 0.97	0.97 / 0.97	0.97
	v.2	0.97 / 0.97	0.97 / 0.97	0.97 / 0.97	0.97
	v.3	0.97 / 0.97	0.97 / 0.97	0.97 / 0.97	0.97

In the case of neural models, both showed consistent performance across all test sets, and none of the data perturbations led to changes in the final scores - all metrics for both models remained exactly the same across all three types of test setups. The **XLM-RoBERTa-Large-En-Ru-MNLI** model ranked last among all models based on precision and recall, with both metrics consistently at 0.97 for both classes — unlike other models, which showed better performance on one class or the other. However, both the F1 and accuracy scores are comparable to those of the classical machine learning models. **ruRoberta-large** performed the best across all metrics, particularly in terms of F1-score and accuracy, where it achieved the highest score of 0.98 for both classes among all models. These results show that both classical and transformer-based models prove to be robust to our attacks, with minimal variation across the test sets.

A consistent trend across classical models is that AIGTs are more frequently misclassified than human-written texts (Figures 7, 8, and 9), while this trend is reversed for ruRoberta-large, which tends to misclassify more human-written texts and misclassified only two AIGTs, 8 and 2 misclassifications, respectively. The only exception to both of these observations is XLM-RoBERTa-Large-En-Ru-MNLI, which showed a nearly balanced distribution of errors between the two classes, 10 and 9 misclassifications, respectively (Figures 11 and 12). This model was also the only one to misclassify texts in the poems genre, both AI-generated and human-written, among all four models investigated in this study.



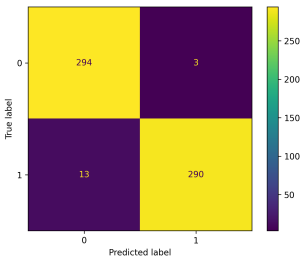
(a) Logistic Regression.



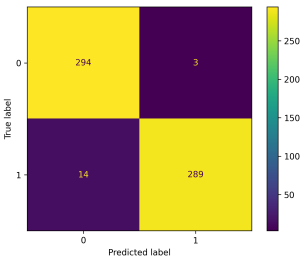
(b) Linear SVM.

Figure 7: TF-IDF: Errors by source and genre for each class in Test Set v.1.

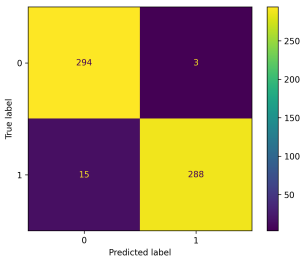
The majority of misclassified texts by both classical models belong to the social media genre—12 AI-generated and 3 human-authored texts for Logistic Regression, and 11 AI-generated and 3 human-written texts for Linear SVM. Two out of three misclassified human texts of social genre belong to the Pikabu subset. In addition, Logistic Regression misclassified one AI-generated news article, while Linear SVM misclassified a news article by Meduza in the v.1 setup. The same AI-generated news article was misclassified again, this time as its 'mixed' and 'n-grams' versions by both models. Beyond that, Logistic Regression misclassified one additional human-edited text, one text humanized with Decopy AI, and one 'n-grams' text in setups v.2 and v.3. Linear SVM misclassified the same text humanized with Decopy AI and 'n-grams' text as Logistic Regression.



(a) Test Set v.1.



(b) Test Set v.2.



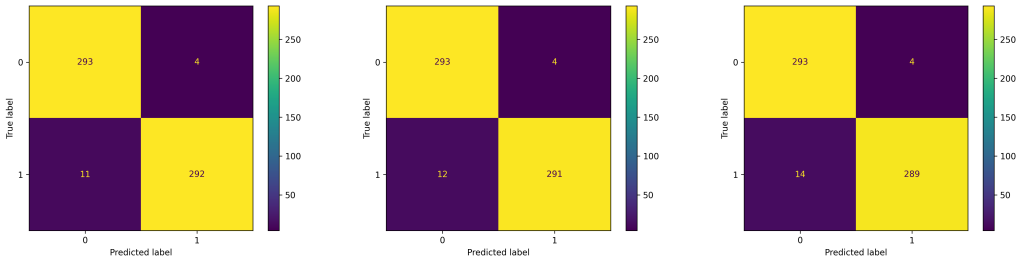
(c) Test Set v.3.

Figure 8: Logistic Regression: Confusion Matrices.

Unlike the classical models, which predominantly misclassified AIGTs, the ruRoberta model struggled more with the classification of human-written texts (Figure 10). In contrast, XLM-RoBERTa-Large-En-Ru-MNLI's errors were more evenly distributed across both classes (Figure 11).

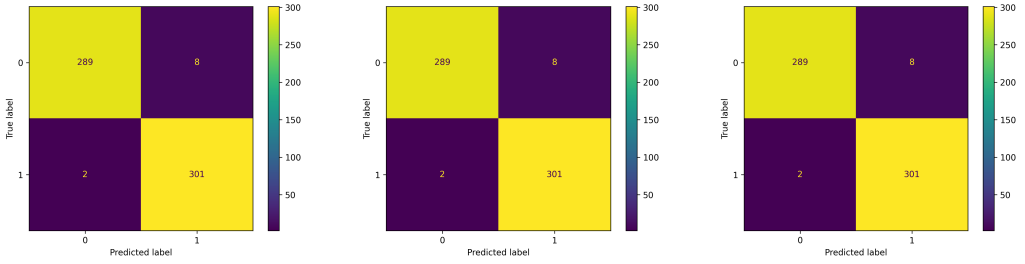
A closer look at genre-level performance reveals additional nuances: while XLM-RoBERTa-Large-En-Ru-MNLI misclassified texts across all three genres - two news articles, eight poems and nine texts of social media genre, all ten errors made by ruRoberta were limited to the social media genre (Figure 12).

The majority of errors in the social media genre for both models—2 out of 3 texts for XLM-RoBERTa-Large-En-Ru-MNLI and 7 out of 8 for ruRoberta—come from the Pikabu dataset (Figure 13), the trend we had previously observed in classical models.



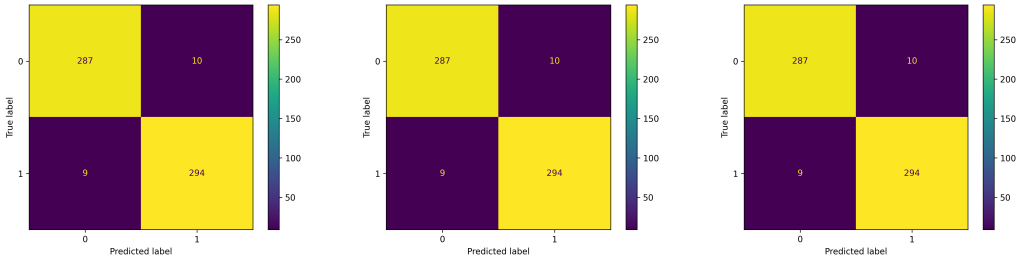
(a) Test Set v.1. (b) Test Set v.2. (c) Test Set v.3.

Figure 9: Linear SVM: Confusion Matrices.



(a) Test Set v.1. (b) Test Set v.2. (c) Test Set v.3.

Figure 10: ruRoberta-large: Confusion Matrices.



(a) Test Set v.1. (b) Test Set v.2. (c) Test Set v.3.

Figure 11: XLM-RoBERTa-Large-En-Ru-MNLI: Confusion Matrices.

The remaining misclassified social media texts come from Vkontakte and Facebook, respectively.

As for the remaining errors by XLM-RoBERTa-Large-En-Ru-MNLI in the news and poems genres, there is no noticeable trend. The mistakes are evenly distributed between authors and sources—one news articles by Lenta and one news article by Ria Novosti, and five poems by different human authors (Figure 14).

Closer review shows that all four models misclassified the following **Pikabu text**: "Ха, а позади меня сидели школотроны, хрустели чипсами и ржали. Суки." as well as the following **AIGT**: "Вот почему некоторых явно коробит от Макса Франка - в ЕР взяли сразу, а в их любимую КПРФ не взяли! Теперь злость сдерживать не

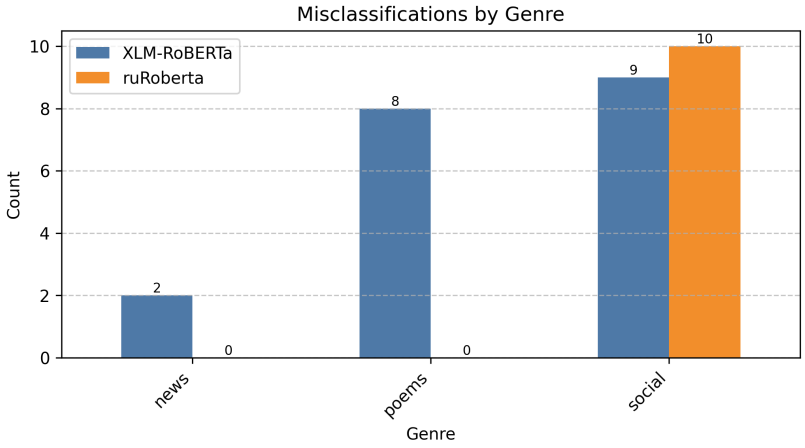
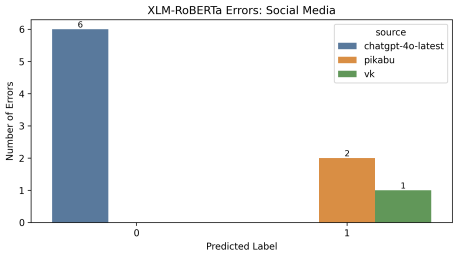
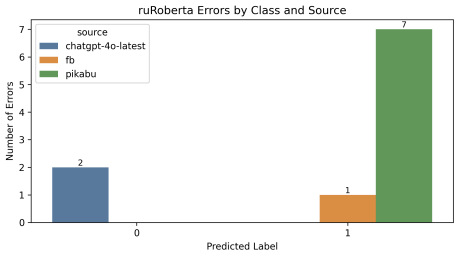


Figure 12: BERT Models: Errors breakdown by genre.

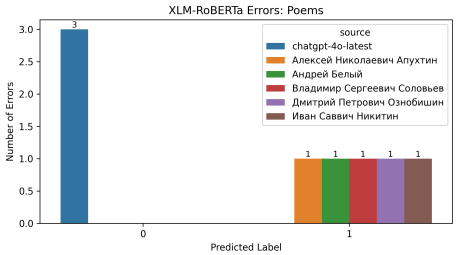


(a) Errors by XLM-RoBERTa.

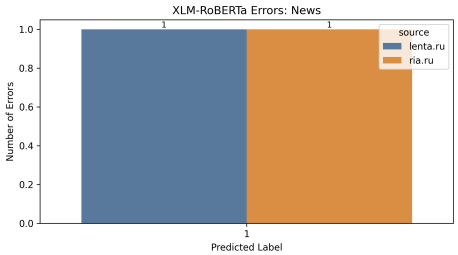


(b) Errors by ruRoberta.

Figure 13: BERT Models: Error distribution in the social media genre by class and source



(a) Poems.



(b) News.

Figure 14: Distribution of XLM-RoBERTa errors across news and poems sources

могут, вопят про капитализм и почему у других лучше. Да просто по характеру не подошел!)". These texts are similar in tone (sarcastic, emotional), style (informal, colloquial, and slang-heavy), and function (social critique or venting)—typical of the kind of content often found in platforms like Pikabu, VK, or forums where users express unfiltered personal reactions. Additionally, both transformer models misclassified another

Pikabu text: "Огромное спасибо за ликбез. Я находила в интернете только часть этого положения.". Furthermore, the following **AI social media post** was misclassified by both classical models and XLM-Roberta: "Купить кофе за 200 рублей: нормально. Купить пирожок за 80 рублей: нормально. Попросить скинуться на общий подарок по 150 рублей: нормально. Заплатить за воду в подъезде 20 рублей: возмущение вселенских масштабов, 'не пользуюсь, почему платить должен'".

There is also a noteworthy overlap in errors made by the classical models. Firstly, both misclassified the exact same set of human-written social media texts. This includes additional texts beyond those already mentioned above that were also misclassified by the transformer models, such as "куртка норм, можно на футболку носить, а вот валенки мы точно зря взяли, резиновые сапоги или осенние носит исключительно этой зимой" from **Pikabu** and "100 лет назад сначала народовольцы и эсеры, потом Цусима, потом Брусиловский прорыв, потом брестский мир, потом красный террор, нэпманы с цыганами, потом гулаг, потом война, потом Гагарин, потом застой, потом бандиты, потом церковь, потом заживем. Имхо" from **Vkotakte**. Secondly, these two models also exhibited similar error patterns in the social media genre for AIGT. Linear SVM misclassified 11 texts, 10 of which were also misclassified by Logistic Regression. Logistic Regression, in turn, made 12 errors in total. This substantial overlap suggests that both models struggle with the same challenging examples in this genre.

6. Discussion

Our experiments reveal that, contrary to our initial expectations (H1), both classical machine learning models and pre-trained transformers demonstrated a high degree of robustness against all three adversarial evasion strategies designed in this study. These findings suggest that, contrary to assumptions underlying our initial hypotheses and some prior studies emphasizing the vulnerability of AIGT detection systems to paraphrasing and lexical substitution attacks (Zhang et al. 2024a; Krishna et al. 2023; Masrour, Emi, and Spero 2025), such weaknesses may be less easily exploited than previously assumed. Although this outcome is encouraging, it highlights the need for deeper research to better understand how detection models operate internally and to identify which aspects of their decision-making processes are genuinely susceptible to manipulation.

The results also call into question our second hypothesis (H2), which predicted that unigram-based manipulation would disrupt classical models. While it was expected that neural models would be less susceptible to this type of attack, the minimal impact observed on classical models was surprising. Despite targeted efforts to remove the most predictive unigrams associated with AIGT and replace them with those indicative of human authorship, classifier performance remained largely unaffected. At present, we can only speculate that, contrary to our initial assumption, the models may not rely solely on isolated lexical cues, or that the manipulated unigrams were not sufficiently influential within their surrounding context to affect classification outcomes. This outcome suggests a need for future work to identify which textual features—beyond unigrams—most strongly influence detection outcomes. Exploring whether more linguistically informed or semantically nuanced substitution strategies can more effectively evade detection remains an open and important direction.

Our third hypothesis (H3) assumed that AI humanizer tools would have only a limited effect on classifier performance in the Russian language setting, due to the lack of research and tool development for languages other than English. This hypothesis is supported by our findings: the generated outputs in Russian often exhibited de-

graded quality, including grammatical inconsistencies, stylistic mismatches, or unnatural phrasing, making them unusable in practical settings. This contrasts with prior research showing that certain humanizer tools can be effective in English (Hua and Yao 2024; Ayub et al. 2024; Masrour, Emi, and Spero 2025), and underscores the importance of language-specific experiments and evaluation.

The strong performance of ruRoberta, a BERT-based model, as the top-performing classifier in our experiments is consistent with previous research that demonstrated the effectiveness of transformer-based approaches in the RuATD task (Posokhov and Makhnytkina 2022; Shamardina et al. 2022; Gritsay, Grabovoy, and Chekhovich 2022), and supports our hypothesis (H4) that BERT-based models would outperform classical methods. However, classical models such as Logistic Regression and Linear SVM closely followed—at times performing on par with— XLM-RoBERTa, challenging the notion that neural models always offer a substantial advantage over simpler baselines. Notably, the strong showing of classical models also highlights their value in scenarios where interpretability is critical, as these models often employ more transparent and explainable decision-making processes compared to the complex and less interpretable nature of deep neural networks.

In addition to our hypothesis-driven findings, several noteworthy observations derived from the error patterns in our experiments. The misclassifications concentrated mainly in the social media genre—especially those from the Pikabu dataset— and suggest genre-specific challenges. Social media texts tend to be informal, stylistically diverse, and noisy, which may obscure the linguistic cues detectors rely on. In contrast, news and poetry genres yielded fewer errors across all models. This highlights the importance of genre diversity in training and evaluating AIGT detectors.

6.1 Limitations and Future Work

This study has several limitations. Firstly, the size of the dataset, though balanced and genre-diverse, remains relatively small for large-scale generalization. Including more genres, topics, and sources, would make it more representative of real-world language use. Additionally, expanding the variety of models used for text generation could enable new experimental setups—such as multi-task classification, and help identify the most effective generation models for Russian. For example, incorporating DeepSeek-V3-Chat[32], which surpassed GPT-4o on the Hugging Face Russian leader board after we completed the generation step, as well as including GigaChat[33] by Sber Devices, a Russian alternative to OpenAI's ChatGPT, would allow for a more comprehensive model comparison. These efforts would not only enrich the scope of this research but would also contribute to bridging the research gap between Russian and high-resource languages like English, where significantly more tools and datasets are currently available (Posokhov and Makhnytkina 2022).

Secondly, there are important limitations associated with the human editing of AIGT. Human edits are inherently subjective, as they rely on human's perception of what constitutes more 'natural' or 'human-like' writing, a process known to be unreliable based on the previous research that shows that humans ultimately cannot confidently recognize AIGT (Liu et al. 2024; Clark et al. 2021). That said, our approach to human editing of AIGT closely mimics a realistic scenario in which a person subtly

32 https://deepseekv3.tech/
33 https://giga.chat/

edits AI-generated content before publishing it (Kashtan and Kipnis 2024). For future research, it would be valuable to investigate the minimum proportion of human edits required to successfully deceive detection systems. Understanding this threshold could contribute to the development of more resilient detectors. It is also worth noting that in many evasion strategies, adversaries intentionally introduce noise or errors to mislead classifiers (Odri and Yoon 2023). However, in our case, we corrected some stylistic and grammatical issues in AIGTs, since the original output was not always of high quality. Although this may have inadvertently reduced the effectiveness of the evasion, it aligns with the goal of simulating a realistic human editing process, where the intent is often to improve rather than to degrade the text.

A three-class classification setup—distinguishing human, AI, and mixed texts—could be explored as a promising direction. Provided that more mixed data becomes available, classifiers could be explicitly trained on such examples, as demonstrated in the study by Zhang et al. (2024a) on English. However, it must also be acknowledged that human editing is a time-consuming and resource-intensive process, making it difficult to apply at scale.

This brings us to the next experimental setup we explored in this study - AI humanizer tools. While many AI humanizer tools were reviewed, only a handful could be fully tested due to limitations in free access, support for Russian, or technical reliability. These constraints may have impacted the depth and scope of the evaluation. Moreover, current research into the impact of AI humanizer tools on AIGT detection remains very limited, even for English. This makes it an important and underexplored area for future investigation. Finally, the study does not include a direct comparison with similar work in English or other high-resource languages. Such cross-linguistic comparisons could help validate whether observed trends hold more broadly or are language-specific.

7. Conclusion

In response to the growing demand for developing reliable AIGT detection methods, this study explored the robustness of baseline classifiers—both classical and neural—against three adversarial evasion strategies: human editing, the use of AI humanizer tools, and targeted manipulation of class-indicative unigrams. To support this investigation, we introduced a new dataset RuMix for Russian AIGT detection task that includes both standard train and test sets, as well as two additional adversarial test sets. Our main contributions include: (1) an extensive overview of AI humanizer tools for Russian; (2) the design and implementation of a suite of adversarial evasion strategies; and (3) the creation and open-source publication of RuMix.

Our findings demonstrate that all models were resilient to the proposed attacks. Neither the humanizer-based nor the unigram perturbations strategies considerably degraded classifier performance, pointing to the limited effectiveness of basic text manipulation strategies. Human editing also proved ineffective in concealing machine authorship—likely due to the well-documented difficulty humans face in reliably discriminating AI-generated text. The BERT-based ruRoberta model achieved the highest overall performance, as expected; however, classical models such as Logistic Regression and Linear SVM also performed competitively. This challenges the assumption that neural methods consistently outperform simpler baselines in this domain.

Overall, our results suggest that while current baseline AIGT detection models are resilient to both basic and several context-aware manipulation strategies tested in this study, important questions remain about their vulnerability to more subtle or systematically optimized attacks. This points to the need for future research focused on

more sophisticated attack strategies, as well as diagnostic tools that can reveal deeper vulnerabilities—especially in underrepresented language settings such as Russian. As an exploratory study, this work offers initial insights into the robustness of current detection models, but more in-depth research is needed to confirm and expand upon these findings. We hope that this work will contribute to ongoing efforts and encourage further research on AIGT detection in the Russian language, where dedicated resources and studies remain limited.

Appendix A: Prompt Template for AI Text Generation

The following template was used as a base and adjusted according to the requirements of each genre.

> You are a [role, e.g., Russian poet, journalist, social media user]. I will provide you with some examples of [text type, e.g., poems, news articles, social media posts] in a moment. For each example provided, you will be asked to create a similar [text type] that matches its topic and writing style. Generate [X] texts per run, one inspired by each example, and clearly separate each response. Only respond with the [text type], and say nothing else. Each response should be approximately [mean word count] words.

Appendix B: List of Evaluated AI Humanizers

Tools Identified in Academic Literature

Tool Name	Subscription	Free Trial	Word Limit	Russian	Text Quality
aihumanizer.ai	Yes	Yes	125	Yes	Poor
Twixify	Yes	Yes	-	No	-
Grammarly Paraphraser	Yes	Yes	125	No	-
Hix AI	Yes	Yes	125	Paid only	-
humanizeai.pro	Yes	Yes	-	Yes	Poor
QuillBot Paraphraser	Yes	Yes	125	Yes	Poor
QuillBot Humanizer	Yes	Yes	125	No	-
GPTinf	Yes	Yes	-	Yes	Moderate
SemihumanAI	Yes	Yes	250/month	Yes	Moderate

Table 2: Overview of AI Humanizer Tools in research papers.

Tools Identified in Online Articles[34][35]

Tool Name	Subscription	Free Trial	Word Limit	Russian	Text Quality
StealthGPT	Yes	No	-	Yes	-
WriteHuman AI	Yes	Yes	200	Yes	Poor
UndetectableAI	Yes	Yes	Few credits only	Yes	Poor
StealthWriter	Yes	Yes	-	Yes	Poor

Table 3: Tools ranked high in online comparison articles.

Based on search in Yandex

Tool Name	Subscription	Free Trial	Word Limit	Russian	Text Quality
YesChatAI Humanizer	Yes	Yes	No	Yes	Good
Decopy.ai	Yes	Yes	50000 char	Yes	Good
humanizeai.org	No	Yes	-	Yes	-

Table 4: AI Humanizer tools appearing first when quering in Russian in Yandex browser.

34 Doha Kash, *Best AI Humanizers For 2025: I Tested 16 Tools, And Only 2 Passed My Test (With Proof)*, Medium, March 11, 2025. Available at: medium.com/@dohakash/...

35 Jean-Marc Buchert, *10 Best AI Humanizer Tools Ranked by Bypass Rate*, Intellectualead, February 28, 2025. Available at: intellectualead.com/...

Appendix C: Extracted Unigrams

Logistic Regression

Top indicative features for AI-generated (class = 1): [’как’ ’где’ ’кто’ ’всё’ ’почему’ ’свет’ ’теперь’ ’ведь’ ’потом’ ’уже’ ’ну’ ’снова’ ’зато’ ’каждый’ ’будто’ ’опять’ ’про’ ’стоит’ ’обещания’ ’новый’ ’такие’ ’тени’ ’ветер’ ’словно’ ’шепчет’ ’ветра’ ’никто’ ’что’ ’очередной’ ’просто’ ’звезды’ ’нас’ ’только’ ’кажется’ ’однако’ ’лица’ ’вчера’ ’продолжают’ ’чтобы’ ’хоть’ ’итоге’ ’новые’ ’путь’ ’напомним’ ’ещё’ ’вдали’ ’цирк’ ’вечный’ ’чайник’ ’конечно’ ’через’ ’лишь’ ’но’ ’толку’ ’могли’ ’когда’ ’интересно’ ’жить’ ’эксперты’ ’пока’]

Top indicative features for Human (class = 0): [’ер’ ’не’ ’при’ ’кпрф’ ’россии’ ’том’ ’есть’ ’он’ ’года’ ’ее’ ’мне’ ’риа’ ’лдпр’ ’много’ ’от’ ’они’ ’было’ ’жириновский’ ’поэтому’ ’единая’ ’коронавируса’ ’ты’ ’едро’ ’после’ ’жирик’ ’надо’ ’очень’ ’партии”сказал’ ’рф’ ’во’ ’стопкоронавирус’ ’первый’ ’россию’ ’того’ ’парнас’ ’ль’ ’сша’ ’именно’ ’меня’ ’единой’ ’моей’ ’человек’ ’числе’ ’выше’ ’августа’ ’11’ ’быть’ ’тебя’ ’президента’ ’грудь’ ’вас’ ’был’ ’марта’ ’россия’ ’зюганов’ ’партию’ ’большинство’ ’следует’ ’почти’]

Linear SVM

Top indicative features for AI-generated (class = 1): [’как’ ’почему’ ’где’ ’кто’ ’приложить’ ’картинку’ ’никто’ ’чайник’ ’потом’ ’могли’ ’ведь’ ’теперь’ ’снова’ ’1800’ ’вт’ ’подсветкой’ ’стоит’ ’всё’ ’каждый’ ’свет’ ’зато’ ’цирк’ ’опять’ ’ну’ ’обещания’ ’кажется’ ’такие’ ’будто’ ’про’ ’хоть’ ’новый’ ’лица’ ’наших’ ’достоевского’ ’новосиба’ ’фанат’ ’шашлыка’ ’лосося’ ’вчера’ ’магазин’ ’ветер’ ’очередной’ ’уже’ ’толку’ ’чтобы’ ’итоге’ ’или’ ’словно’ ’уровень’ ’просто’ ’недавно’ ’шепчет’ ’через’ ’продолжают’ ’жить’ ’думаешь’ ’делает’ ’новые’ ’бумаги’ ’цены’]

Top indicative features for Human (class = 0): [’ер’ ’не’ ’кпрф’ ’россии’ ’при’ ’лдпр’ ’есть’ ’он’ ’едро’ ’жириновский’ ’жирик’ ’том’ ’единая’ ’много’ ’поэтому’ ’именно’ ’ее’ ’мне’ ’было’ ’они’ ’надо’ ’парнас’ ’года’ ’россию’ ’от’ ’очень’ ’жирика’ ’владимир’ ’того’ ’путин’ ’тролль’ ’долбоёб’ ’вольфович’ ’первый’ ’ты’ ’дмитрий’ ’пенсион’ ’жириновским’ ’единой’ ’выше’ ’почти’ ’во’ ’единую’ ’работает’ ’рф’ ’сказал’ ’риа’ ’после’ ’довольно’ ’ль’ ’большое’ ’тебя’ ’большинство’ ’депутаты’ ’кем’ ’партию’ ’вас’ ’аватарку’ ’маразм’ ’поменять’]

References

Adelani, David Ifeoluwa, Haotian Mai, Fuming Fang, Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. 2019. Generating sentiment-preserving fake online reviews using neural language models and their human- and machine-based detection.

Alemohammad, Sina, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. 2023. Self-consuming generative models go mad.

Andreev, Nikita, Alexander Shirnin, Vladislav Mikhailov, and Ekaterina Artemova. 2024. Papilusion at dagpap24: Paper or illusion? detecting ai-generated scientific papers.

Ayub, T., R. Ahmad Malla, M. Y. Khan, and S. A. Ganaie. 2024. The art of deception: humanizing ai to outsmart detection. *Global Knowledge, Memory and Communication*, ahead-of-print(ahead-of-print).

Bearman, Margaret, Janice Ryan, and Rola Ajjawi. 2023. Discourses of artificial intelligence in higher education: a critical literature review. *Higher Education*, 86:369–385.

Besnier, Victor, Himalaya Jain, Andrei Bursuc, Matthieu Cord, and Patrick Pérez. 2020. This dataset does not exist: Training models from generated images. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Cai, Shuyang and Wanyun Cui. 2023. Evade chatgpt detectors via a single space.

Chamezopoulos, Savvas, Drahomira Herrmannova, Anita De Waard, Drahomira Herrmannova, Domenic Rosati, and Yury Kashnitsky. 2024. Overview of the DagPap24 shared task on detecting automatically generated scientific paper. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 7–11, Association for Computational Linguistics, Bangkok, Thailand.

Chen, Yutian, Hao Kang, Vivian Zhai, Liangze Li, Rita Singh, and Bhiksha Raj. 2023. Gpt-sentinel: Distinguishing human and chatgpt generated content.

Clark, Elizabeth, 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.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Dai, Long, Jiarong Mao, Xuefeng Fan, and Xiaoyi Zhou. 2022. Deephider: A covert nlp watermarking framework based on multi-task learning.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Fomenko, Vadim. 2020. Russian news 2020. https://www.kaggle.com/datasets/vfomenko/russian-news-2020. Accessed: 2025-01-06.

Fraser, Kathleen C., Hillary Dawkins, and Svetlana Kiritchenko. 2024. Detecting ai-generated text: Factors influencing detectability with current methods.

Fröhling, Leon and Arkaitz Zubiaga. 2021. Feature-based detection of automated language models: tackling gpt-2, gpt-3 and grover. *PeerJ Computer Science*, 7.

Gao, Christopher A., Fiona M. Howard, Nikolay S. Markov, et al. 2023. Comparing scientific abstracts generated by chatgpt to real abstracts with detectors and blinded human reviewers. *npj Digital Medicine*, 6:75.

Gowal, Sven, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, and Timothy Mann. 2021. Improving robustness using generated data.

Gritsai, German, Ildar Khabutdinov, and Andrey Grabovoy. 2024. Multi-head span-based detector for ai-generated fragments in scientific papers. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, page 220–225, Association for Computational Linguistics.

Gritsai, German, Anastasia Voznyuk, Andrey Grabovoy, and Yury Chekhovich. 2025. Are ai detectors good enough? a survey on quality of datasets with machine-generated texts.

Gritsay, German, Andrey Grabovoy, and Yury Chekhovich. 2022. Automatic detection of machine generated texts: Need more tokens. In *2022 Ivannikov Memorial Workshop (IVMEM)*, pages 20–26.

Hu, Xiaomeng, Pin-Yu Chen, and Tsung-Yi Ho. 2023. Radar: Robust ai-text detection via adversarial learning.

Hua, Heng and Chia-Jung Yao. 2024. Investigating generative ai models and detection techniques: impacts of tokenization and dataset size on identification of ai-generated text.

Frontiers in Artificial Intelligence, 7:1469197.
Ilya Gusev. 2024. pikabu (revision 96466c2).
Jawahar, Ganesh, Muhammad Abdul-Mageed, and Laks Lakshmanan, V.S. 2020a. Automatic detection of machine generated text: A critical survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2296–2309, International Committee on Computational Linguistics, Barcelona, Spain (Online).
Jawahar, Ganesh, Muhammad Abdul-Mageed, and Laks V. S. Lakshmanan. 2020b. Automatic detection of machine generated text: A critical survey.
Jordon, James, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N. Cohen, and Adrian Weller. 2022. Synthetic data – what, why and how?
Kashtan, Idan and Alon Kipnis. 2024. An information-theoretic approach for detecting edits in ai-generated text. *Harvard Data Science Review*, (Special Issue 5).
Kreps, Sarah, R. Miles McCain, and Miles Brundage. 2022. All the news that's fit to fabricate: Ai-generated text as a tool of media misinformation. *Journal of Experimental Political Science*, 9(1):104–117.
Krishna, Kalpesh, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense.
Kuditipudi, Rohith, John Thickstun, Tatsunori Hashimoto, and Percy Liang. 2024. Robust distortion-free watermarks for language models.
Liang, Weixin, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel A. McFarland, and James Y. Zou. 2024. Monitoring ai-modified content at scale: A case study on the impact of chatgpt on ai conference peer reviews.
Liang, Weixin, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. Gpt detectors are biased against non-native english writers.
Liu, Yinhan, 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.
Liu, Zeyan, Zijun Yao, Fengjun Li, and Bo Luo. 2024. On the detectability of chatgpt content: Benchmarking, methodology, and evaluation through the lens of academic writing.
Lu, Ning, Shengcai Liu, Rui He, Qi Wang, Yew-Soon Ong, and Ke Tang. 2024. Large language models can be guided to evade ai-generated text detection.
Maloyan, Narek, Bulat Nutfullin, and Eugene Ilyshin. 2022. Dialog-22 ruatd generated text detection. In *Computational Linguistics and Intellectual Technologies*, page 394–401, RSUH.
Masrour, Elyas, Bradley Emi, and Max Spero. 2025. Damage: Detecting adversarially modified ai generated text.
Messeri, Lisa and Molly J. Crockett. 2024. Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627:49–58.
Odri, Guillaume-Anthony and Diane Ji Yun Yoon. 2023. Detecting generative artificial intelligence in scientific articles: Evasion techniques and implications for scientific integrity. *Orthopaedics & Traumatology: Surgery & Research*, 109(8):103706.
Orzhenovskii, Mikhail. 2022. Detecting auto-generated texts with language model and attacking the detector. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2022"*, pages 412–419, Moscow, Russia.
Posokhov, Pavel and Olesia Makhnytkina. 2022. Artificial text detection in russian language: A bert-based approach. In *Proceedings of the Conference*, pages 470–476.
Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.
Sadasivan, Vinu Sankar, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2025. Can ai-generated text be reliably detected?
Shamardina, Tatiana, Vladislav Mikhailov, Daniil Chernianskii, Alena Fenogenova, Marat Saidov, Anastasiya Valeeva, Tatiana Shavrina, Ivan Smurov, Elena Tutubalina, and Ekaterina Artemova. 2022. Findings of the the ruatd shared task 2022 on artificial text detection in russian. In *Computational Linguistics and Intellectual Technologies*, page 497–511, RSUH.
Shamardina, Tatiana, Marat Saidov, Alena Fenogenova, Aleksandr Tumanov, Alina Zemlyakova, Anna Lebedeva, Ekaterina Gryaznova, Tatiana Shavrina, Vladislav Mikhailov, and Ekaterina Artemova. 2024. Coat: Corpus of artificial texts. *Natural Language Processing*, 31:1–26.

Shavrina, Tatiana and Olga Shapovalova. 2017. To the methodology of corpus construction for machine learning: «taiga» syntax tree corpus and parser. In *Proceedings of CORPORA 2017 International Conference*, Saint Petersburg, Russia.

Shumailov, Ilia, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. Ai models collapse when trained on recursively generated data. *Nature*, 631:755–759.

Silkin, Georgy. n.d. 19,000+ russian poems. https://www.kaggle.com/datasets/grafstor/19-000-russian-poems/data. Accessed: 2025-01-06.

Wang, Changhan, Kyunghyun Cho, and Jiatao Gu. 2019. Neural machine translation with byte-level subwords.

Wang, Yuxia, Jonibek Mansurov, Petar Ivanov, jinyan su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Chenxi Whitehouse, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024a. Semeval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2041–2063, Association for Computational Linguistics, Mexico City, Mexico.

Wang, Yuxia, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, Thomas Arnold, Alham Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024b. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1369–1407, Association for Computational Linguistics, St. Julian's, Malta.

Weidinger, Laura, 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. 2021. Ethical and social risks of harm from language models.

Yang, Yiben, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. 2020. Generative data augmentation for commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1008–1025, Association for Computational Linguistics, Online.

Zellers, Rowan, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2020. Defending against neural fake news.

Zhang, Lingze and Ellie Pavlick. 2025. Does training on synthetic data make models less robust? ArXiv preprint arXiv:2502.07164, version 1.

Zhang, Qihui, Chujie Gao, Dongping Chen, Yue Huang, Yixin Huang, Zhenyang Sun, Shilin Zhang, Weiye Li, Zhengyan Fu, Yao Wan, and Lichao Sun. 2024a. Llm-as-a-coauthor: Can mixed human-written and machine-generated text be detected?

Zhang, Ye, Qian Leng, Mengran Zhu, Rui Ding, Yue Wu, Jintong Song, and Yulu Gong. 2024b. Enhancing text authenticity: A novel hybrid approach for ai-generated text detection.

Zhang, Yizhou, Yicheng Ma, Jiaxin Liu, Xiaohan Liu, Xiaodong Wang, and Wei Lu. 2024c. Detection vs. anti-detection: Is text generated by ai detectable? In *Wisdom, Well-Being, Win-Win. iConference 2024*, volume 14596 of *Lecture Notes in Computer Science*, Springer, Cham.

Zhao, Yuan, Junruo Gao, Junlin Wang, Gang Luo, and Liang Tang. 2024. Utilizing an ensemble model with anomalous label smoothing to detect generated scientific papers. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 130–134, Association for Computational Linguistics, Bangkok, Thailand.

Zmitrovich, Dmitry, Alexander Abramov, Andrey Kalmykov, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Tatiana Shavrina, Sergey Markov, Vladislav Mikhailov, and Alena Fenogenova. 2023. A family of pretrained transformer language models for russian.