T. Batura, E. Bruches, M. Shvenk, V. Malykh

AINL-EVAL 2025 SHARED TASK: DETECTION OF AI-GENERATED SCIENTIFIC ABSTRACTS IN RUSSIAN

ABSTRACT. The rapid advancement of large language models (LLMs) has revolutionized text generation, making it increasingly difficult to distinguish between human- and AI-generated content. This poses a significant challenge to academic integrity, particularly in scientific publishing and multilingual contexts where detection resources are often limited. To address this critical gap, we introduce the AINL-Eval 2025 Shared Task, specifically focused on the detection of AI-generated scientific abstracts in Russian. We present a novel, large-scale dataset comprising 52,305 samples, including human-written abstracts across 12 diverse scientific domains and AI-generated counterparts from five state-of-the-art LLMs (GPT-4-Turbo, Gemma2-27B, Llama3.3-70B, Deepseek-V3, and GigaChat-Lite). A core objective of the task is to challenge participants to develop robust solutions capable of generalizing to both (i) previously unseen scientific domains and (ii) models not included in the training data. The task was organized in two phases, attracting 10 teams and 159 submissions, with top systems demonstrating strong performance in identifying AI-generated content. We also establish a continuous shared task platform to foster ongoing research and long-term progress in this important area. The dataset and platform are publicly available at: `https://github.com/iis-research-team/AINL-Eval-2025`

§1. INTRODUCTION

In recent years, the development of large language models (LLMs) has revolutionized natural language processing. These models are now capable of generating text that closely resembles human writing, making it increasingly difficult to distinguish between AI and human-generated content. This capability has led to numerous applications across various domains, with researchers proposing LLM-generated paper texts [1, 2], academic posters [3], and even research ideas [4–6]. Although this represents

significant technological progress, there are domains in which the unrestricted use of LLM raises concerns, particularly in scientific publishing and academic integrity. In response to these challenges, several detection tools have emerged, such as LLM-DetectAIve [7] and M4 [8], designed to identify AI-generated content.

The scientific community faces a growing challenge as AI-generated papers become more sophisticated and harder to detect. This is particularly concerning in multilingual contexts where detection tools and evaluation benchmarks may be less developed for languages other than English. To address this gap, we introduce AINL-Eval 2025, a shared task focused on detecting AI-generated scientific abstracts in Russian. Unlike previous efforts such as the RuATD Shared Task 2022 [9], which addressed AI-generated texts across multiple domains, including machine translation and paraphrase generation, our task focuses specifically on scientific texts in Russian, creating a specialized testbed for academic content integrity.

In this paper, we introduce a new dataset of scientific abstracts in Russian to distinguish between human- and AI-generated texts and design challenges that require participants to develop solutions capable of generalizing to new domains and detecting texts generated by models not included in the training data. Additionally, we have created a continuous shared task platform that remains accessible for community contributions and supports long-term progress in this important area.

§2. Related Work

Generally, methods for detecting AI-generated text fall into three categories: watermarking techniques, statistical approaches, and machine learning-based methods.

Watermarking is a technique designed to incorporate robust detection signals into machine-generated text [10]. A watermarking method typically consists of three components: watermark, encoder, and decoder. An encoder E embeds a watermark into a content, while a decoder D decodes a watermark from a content (watermarked or unwatermarked). When a content has the watermark w, the decoded watermark is similar to w. Watermarking can be implemented using hand-crafted heuristics [11, 12] or using neural networks-based methods [13, 14]. Such methods are useful if the model is known.

Statistical methods (including stylistic analysis) rely on features extracted from the text. For example, in [15] the features from 6 categories

are taken into account, e.g. lexical features (word count, char count, etc.) or named entity (first person count, direct address count, etc.). The study [16] showed that this set of features is highly correlated with our cognitive processes and may be used to distinguish between human-written and AI-generated content. In [17] the authors use the following groups of features to detect AI-generated tweets: 1) Phraseology - features which quantify how the author organizes words and phrases when creating a piece of text (e.g., avg. word, sent. count, etc.), 2) Punctuation – features to quantify how the author utilizes different punctuation (e.g., avg. unique punctuation count) and 3) Linguistic Diversity - features to quantify how the author uses different words in the writing (e.g., richness and readability scores). While statistical methods offer reliability and interpretability, they often struggle with broader applicability due to their reliance on pre-defined feature sets.

Machine learning algorithms do not involve an explicit feature extraction step, as described in the previous sections. The classifier is given the entire text as input and must learn, as part of the training process, which characteristics of the text differ between the classes. In [18] the authors propose a system which is based on XLM-longformer with CRF layer. In [19] XGB-classifier and SVM are applied for this task. LLM-DetectAIve [20] uses fine-tuned RoBERTa and DeBERTa to distinguish between four categories: (i) human-written, (ii) machine-generated, (iii) machine-written, then machine-humanized, and (iv) human-written, then machine-polished.

Recently, a number of **zero-shot methods** were proposed. The main idea is to evaluate the average per-token log probability of the generated text and thresholding [21]. DetectGPT [22] uses a property of the structure of an LLM's probability function. GPT-Who [23] employs the Uniform Information Density (UID) principle, assuming that humans prefer to spread information evenly during language production.

Given the growing significance of this field, numerous **academic competitions** have been established to assess progress. SemEval-2024 Task 8: Multidomain, Multimodel and Multilingual Machine-Generated Text Detection [24] featured three subtasks: (1) Human vs. Machine Classification - the goal of this subtask is to accurately classify a text as either produced by a human or generated by a machine; (2) Multi-Way Generator Detection aims to pinpoint the exact source of a text, i.e., determine whether it originated from a human or a specific LLM; (3) Changing Point Detection - the goal is to precisely identify the exact boundary (changing

point) within a text at which the authorship transitions from a human to machine happens. The RuATD Shared Task 2022 [9] was developed for Russian language and consisted of two sub-tasks: (1) to determine if a given text is automatically generated or written by a human; (2) to identify the author of a given text. This evaluation is designed specifically to detect AI-generated scientific texts in Russian. GenAI Content Detection Task 1: English and Multilingual Machine-Generated Text Detection: AI vs. Human [25] has two subtasks: monolingual (English) and multilingial, where the data comes from more than 8 different domains, e.g. scientific papers, social media, emails, etc. DAGPap22 shared task [26] concentrates on the detection of AI-generated scientific papers. The ALTA shared tasks [27] aims to discriminate between human-written and synthetic text generated by LLM.

§3. Dataset

3.1. Text generation. The overall corpus includes the texts from six generators, i.e., one human writer and five different LLMs.The human-written abstracts were parsed and cleaned from the digital scientific journals in Russian. The domains were the following: Math, Philology, Physics, Chemistry, Pedagogy, IT, Law, Medicine, Oil and Gas, Management, Economics, Biology. It is worth noting that Economics and Biology domains were not presented in train and dev sets but only appears in the testing stage.

For each human-written text, the models were prompted to generate the abstract based on the title and keywords. The prompt in Russian was as follows:

> Given the paper title and keywords,
> generate the abstract for this paper.
> Write only the abstract text. Don't
> use the phrase "In this work" in the
> beginning of the text.
> Title: {title}.
> Keywords: {keywords}

We prompted each model with the same prompt without changing it. The following models were selected to generate abstracts: GPT-4-Turbo [28], Gemma2-27B [29], Llama3.3-70B [30], Deepseek-V3 [31] and GigaChat-Lite [32].

The post-processing stage includes removing the LLM artifacts such as specific prefixes or inappropriate output. Also, we noticed that the models tend to begin the generation with specific patterns, so we implemented some heuristics to change the beginnings while preserving the main content of the abstracts as is.

Thus, we invite participants to propose solutions to the following key challenges:

(1) Handling data that extend beyond the training set (generalization to new domains).

(2) Detecting texts generated by a model not included in the training data (generalization to new models).

3.2. Dataset overview. The dataset size is 52,305 samples, having 35,158 samples for train, 10,978 samples for dev and 6,169 samples for test. The distribution of labels within the subsets is uniform.

The quantitative analysis of the training subset reveals several findings regarding abstract length. Human-written abstracts are significantly longer, averaging 126.4 words. In contrast, model-generated abstracts are considerably shorter, ranging from an average of 49.6 tokens for GigaChat-Lite to 85.7 tokens for GPT-4-Turbo. It is worth noting that IT and Philology are the domains with the longest human-written abstracts. Another interesting finding is that humans, Llama-3.3 and GPT-4-Turbo have the closest to the average number of words in the sentence, while Gemma2-27B and GigaChat-Lite generate sentences with a fewer number of words. Another distinguishing feature is usage of digitals: humans append in 10 times more digits in the texts than the models.

§4. Task organization

The general purpose of the competition is to identify the precise origin of a given text, determining if it was authored by a human or generated by a particular large language model. The texts are abstracts of the scientific papers in Russian.

The shared task was run in two phases.

Development phase. The training and development data were available to the participants. The training data contained the texts and the corresponding labels reflecting the author of the text: human, GPT4-Turbo, Gemma2-27B and Llama3.3-70B. These data are assumed to be used to develop the system. The development data contained additional generations

from the unseen model which was GigaChat-Lite. The participants didn't know the ground truth labels for the development set but they could submit the results on Codalab and get the results. We didn't limit the number of submission during this stage. The leaderboard showed the best results for each participating team, regardless of the submission time.

Test phase. To assess the system's ability to generalization during the test phase both new domains (Economics and Biology) and generator (DeepSeek) were presented. This private set contained only the texts, without ground truth labels. The duration of this stage was one week. The participants had only 5 attempts to submit their results. The number of submissions was limited to avoid overfitting on the test data. The submission with the highest score was considered to be the final team's result.

After the competition ended, we released the gold labels for both the development and test sets. Furthermore, we kept the submission system open for the test dataset for post-shared task evaluations.

§5. Evaluation and results

We received 159 submissions from ten teams in the development stage and five teams in the test stage. Accuracy was used to evaluate the submissions, as in the RuATD Shared Task 2022 [9]. Table 1 shows the scores on the development set for all submitted systems. The top two systems on the development set - GigaCheck (Mistral-7B) and YandexGPT 8B - demonstrated a clear performance advantage over traditional methods and baselines.

Team **sastsy** leveraged Mistral-7B-v0.3 as its backbone LLM, enhanced with a dual-head architecture. The first head performs binary classification (Human vs. AI), while the second identifies specific AI models (e.g., GPT-4, LLaMA-3) through multiclass classification. To optimize training efficiency, the model employs LoRA (Low-Rank Adaptation), enabling lightweight fine-tuning with minimal parameter updates. Additionally, a weighted cross-entropy loss is used to ensure balanced learning across imbalanced datasets, thereby improving detection accuracy. Further implementation details are described in the paper [33].

Team **adugeen** applied a combined approach based on statistical and neural model features to improve overall detection performance. In such a hybrid architecture, the linear layers and the classifier were fine-tuned, while the rest of the model was kept frozen. Fine-tuning the YandexGPT

User	Entries	Accuracy	System Summary
sastsy	29	**0.9122**	GigaCheck (Mistral-7B)
adugeen	64	0.8696	YandexGPT 8B
Nick	13	0.8245	n/a
vikosik3000	3	0.8164	n/a
Baseline	-	*0.8081*	*LogReg + TF-IDF*
kelijah	2	0.8068	n/a
fedrshm	10	0.7999	n/a
Baseline	-	*0.7903*	*BERT*
chrnegor	12	0.7833	n/a
dorj	5	0.7564	n/a
FedorinovVladislav	2	0.6468	n/a
eborisov	1	0.2009	n/a

Table 1. Accuracy on the development set.

8B model yielded the best performance on the development set. On the test set, the best results were achieved using a combination of bag-of-words features and binoculars derived from the Gemma 2B and LLaMA 1B models. This team submitted the highest number of entries to the leaderboard.

Overall, the final results on the test set (see Table 2) further emphasize the growing dominance of large language models in achieving high accuracy in the detection of AI-generated scientific abstracts in Russian.

User	Entries	Accuracy	System Summary
sastsy	3	**0.8635**	GigaCheck (Mistral-7B)
adugeen	4	0.8462	BoW + b-Gemma 2B + b-Llama 1B
vikosik3000	1	0.8159	n/a
Baseline	-	*0.8105*	*LogReg + TF-IDF*
ESBaklanova	1	0.2099	n/a
fedrshm	9	0.2072	n/a

Table 2. Accuracy on the test set.

§6. Limitations

Due to resource limits, we only generated abstracts. The more challenging task is to generate the whole paper text. Another limitation is that the generation was performed based only on the title and keywords. However, using the text or other metadata could improve the generated texts. Also the task was limited by the classification whether the whole text of abstract is generated or not. But in practice, the most general case is to edit the generated text or to generate the more proof-read version of the human texts. The task of AI-generated spans detection is considered as one of the future research direction.

§7. Conclusion

We presented the AINL-Eval 2025 Shared Task, focused on the detection of AI-generated scientific abstracts in Russian. The best solution for the shared task achieved 91.22% accuracy on the development set and 86.35% accuracy on the test set. By introducing a comprehensive, multi-domain, and multi-model dataset, we have provided a specialized testbed for evaluating the robustness and generalizability of AI-generated text detection systems. The task design, which explicitly challenged participants to handle unseen domains and models, pushed the boundaries of current detection capabilities and highlighted the need for more sophisticated and adaptable solutions.

References

1. E. H. Ayache and C. M. B. Omand, *Generating Scientific Articles with Machine Learning*, ArXiv preprint arXiv:2203.16569 (2022).
2. H. Chen, H. Takamura, and H. Nakayama, *SciXGen: A Scientific Paper Dataset for Context-Aware Text Generation*, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Punta Cana, Dominican Republic, Association for Computational Linguistics, 2021, pp. 1483–1492.
3. S. Xu and X. Wan, *Neural Content Extraction for Poster Generation of Scientific Papers*, ArXiv preprint arXiv:2112.08550 (2021).
4. X. Gu and M. Krenn, *Interesting Scientific Idea Generation using Knowledge Graphs and LLMs: Evaluations with 100 Research Group Leaders*, ArXiv preprint arXiv:2405.17044 (2024).
5. M. Radensky, S. Shahid, R. Fok, P. Siangliulue, T. Hope, and D. S. Weld, *Scideator: Human-LLM Scientific Idea Generation Grounded in Research-Paper Facet Recombination*, ArXiv preprint arXiv:2409.14634 (2024).

6. C. Si, D. Yang, and T. Hashimoto, *Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers*, ArXiv preprint arXiv:2409.04109 (2024).

7. M. Abassy, K. Elozeiri, A. Aziz, M. N. Ta, R. V. Tomar, B. Adhikari, S. El Dine Ahmed, Y. Wang, O. Mohammed Afzal, Z. Xie, J. Mansurov, E. Artemova, V. Mikhailov, R. Xing, J. Geng, H. Iqbal, Z. M. Mujahid, T. Mahmoud, A. Tsvigun, A. F. Aji, A. Shelmanov, N. Habash, I. Gurevych, and P. Nakov, *LLM-DetectAIve: a Tool for Fine-Grained Machine-Generated Text Detection*, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Miami, Florida, USA, Association for Computational Linguistics, 2024, pp. 3360–343.

8. Y. Wang, J. Mansurov, P. Ivanov, J. Su, A. Shelmanov, A. Tsvigun, C. Whitehouse, O. Mohammed Afzal, T. Mahmoud, T. Sasaki, T. Arnold, A. Aji, N. Habash, I. Gurevych, and P. Nakov, *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), Association for Computational Linguistics, 2024, pp. 1369–1407.

9. T. Shamardina, V. Mikhailov, D. Chernianskii, A. Fenogenova, M. Saidov, A. Valeeva, T. Shavrina, I. Smurov, E. Tutubalina, and E. Artemova, *Findings of the The RuATD Shared Task 2022 on Artificial Text Detection in Russian*, ArXiv preprint arXiv:2206.01583 (2022).

10. J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein, *A Watermark for Large Language Models*, in: Proceedings of Machine Learning Research, 2023, pp. 17061–17084.

11. N. Bi, Q. Sun, D. Huang, Z. Yang, and J. Huang, *Robust Image Watermarking Based on Multiband Wavelets and Empirical Mode Decomposition. —* Trans. Img. Proc. **16**(8) (2007), 1956–1966.

12. Y. Wen, J. Kirchenbauer, J. Geiping, and T. Goldstein, *Tree-rings Watermarks: Invisible Fingerprints for Diffusion Images*, in: Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23), Curran Associates Inc., Red Hook, NY, USA, Article 2529, 2023, pp. 58047–58063.

13. P. Fernandez, G. Couairon, H. Jegou, M. Douze, and T. Furon, *The Stable Signature: Rooting Watermarks in Latent Diffusion Models*, in: ICCV 2023 — International Conference on Computer Vision, Paris, France, Oct. 2023, 2023, pp. 22409–22420.

14. S. Abdelnabi and M. Fritz, *Adversarial Watermarking Transformer: Towards Tracing Text Provenance with Data Hiding*, 42nd IEEE Symposium on Security and Privacy, 2021, pp. 121–140.

15. C. Opara, *StyloAI: Distinguishing AI-Generated Content with Stylometric Analysis*, in: Proceedings of 25th International Conference on Artificial on Artificial Intelligence in Education (AIED 2024), 2024.

16. C. Opara, *Distinguishing AI-Generated and Human-Written Text Through Psycholinguistic Analysis*, ArXiv preprint arXiv:2505.01800 (2025).

17. T. Kumarage, J. Garland, A. Bhattacharjee, K. Trapeznikov, S. Ruston, and H. Liu, *Stylometric Detection of AI-Generated Text in Twitter Timelines*, ArXiv preprint arXiv:2303.03697 (2023).

18. R. M. R. Kadiyala, S. Pullakhandam, K. Mehreen, D. Sharma, S. Gupta, J. Purbey, A. Srivastava, S. TippaReddy, A. R. Bobbili, S. Telugara Chandrashekhar, M. Adeeb, S. Vura, and H. Farooq, *Robust and Fine-Grained Detection of AI Generated Texts*, ArXiv preprint arXiv:2504.11952 (2025).

19. N. Prova, *Detecting AI Generated Text Based on NLP and Machine Learning Approaches*, ArXiv preprint arXiv:2404.10032 (2024).

20. M. Abassy, K. Elozeiri, A. Aziz, M. N. Ta, R. V. Tomar, B. Adhikari, S. El Dine Ahmed, Y. Wang, O. Mohammed Afzal, Z. Xie, J. Mansurov, E. Artemova, V. Mikhailov, R. Xing, J. Geng, H. Iqbal, Z. M. Mujahid, T. Mahmoud, A. Tsvigun, A. F. Aji, A. Shelmanov, N. Habash, I. Gurevych, and P. Nakov, *LLM-DetectAIve: a Tool for Fine-Grained Machine-Generated Text Detection*, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2024, pp. 336–343.

21. I. Solaiman, M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu, A. Radford, G. Krueger, J. W. Kim, S. Kreps, M. McCain, A. Newhouse, J. Blazakis, K. McGuffie, and J. Wang, *Release Strategies and the Social Impacts of Language Models*, ArXiv preprint arXiv:1908.09203 (2019).

22. E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, *DetectGPT: Zero-Shot Machine-Generated Text Detection Using Probability Curvature*, in: Proceedings of the 40th International Conference on Machine Learning (ICML'23), vol. 202, 2023, pp. 24950–24962.

23. S. Venkatraman, A. Uchendu, and D. Lee, *GPT-who: An Information Density-based Machine-Generated Text Detector*, in: Findings of the Association for Computational Linguistics: NAACL 2024, 2024, pp. 103–115.

24. Y. Wang, J. Mansurov, P. Ivanov, J. Su, A. Shelmanov, A. Tsvigun, O. Mohammed Afzal, T. Mahmoud, G. Puccetti, and T. Arnold, *SemEval-2024 Task 8: Multidomain, Multimodel and Multilingual Machine-Generated Text Detection*, in: Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), 2024, pp. 2057–2079.

25. Y. Wang, A. Shelmanov, J. Mansurov, A. Tsvigun, V. Mikhailov, R. Xing, Z. Xie, J. Geng, G. Puccetti, E. Artemova, J. Su, M. N. Ta, M. Abassy, K. A. Elozeiri, S. El Dine Ahmed El Etter, M. Goloburda, T. Mahmoud, R. V. Tomar, N. Laiyk, O. Mohammed Afzal, R. Koike, M. Kaneko, A. F. Aji, N. Habash, I. Gurevych, and P. Nakov, *GenAI Content Detection Task 1: English and Multilingual Machine-Generated Text Detection: AI vs. Human*, in: Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect), 2025, pp. 244–261.

26. Y. Kashnitsky, D. Herrmannova, A. de Waard, G. Tsatsaronis, C. C. Fennell, and C. Labbe, *Overview of the DAGPap22 Shared Task on Detecting Automatically Generated Scientific Papers*, in: Proceedings of the Third Workshop on Scholarly Document Processing, 2022, pp. 210–213.

27. D. Molla, H. Zhan, X. He, and Q. Xu, *Overview of the 2023 ALTA Shared Task: Discriminate between Human-Written and Machine-Generated Text*, in: Proceedings

of the 21st Annual Workshop of the Australasian Language Technology Association, 2023, pp. 148–152.

28. OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, *et al.*, *GPT-4 Technical Report*, ArXiv preprint arXiv:2303.08774 (2023).

29. M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, Léonard Hussenot, T. Mesnard, B. Shahriari, A. Ramé, J. Ferret, *et al.*, *Gemma 2: Improving Open Language Models at a Practical Size*, ArXiv preprint arXiv:2408.00118 (2024).

30. A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, *et al.*, *The Llama 3 Herd of Models*, ArXiv preprint arXiv:2407.21783 (2024).

31. A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, D. Dai, D. Guo, D. Yang, D. Chen, *et al.*, *DeepSeek-V3 Technical Report*, ArXiv preprint arXiv:2412.19437 (2024).

32. V. Mamedov, E. Kosarev, G. Leleytner, I. Shchuckin, V. Berezovskiy, D. Smirnov, D. Kozlov, *et al.*, *GigaChat Family: Efficient Russian Language Modeling Through Mixture of Experts Architecture*, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics. Volume 3: System Demonstrations, 2025, pp. 93–106.

33. I. Tolstykh, A. Tsybina, S. Yakubson, A. Gordeev, V. Dokholyan, and M. Kuprashevich, *GigaCheck: Detecting LLM-generated Content*, ArXiv preprint arXiv:2410.23728 (2024).

A.P. Ershov Institute
of Informatics Systems,
Novosibirsk, Russia

E-mail: tatiana.v.batura@gmail.com

A.P. Ershov Institute
of Informatics Systems, Novosibirsk;
Novosibirsk State University,
Novosibirsk, Russia

E-mail: bruches@bk.ru

Novosibirsk State University,
Novosibirsk, Russia

E-mail: m.shvenk@g.nsu.ru

ITMO University,
St. Petersburg, Russia

E-mail: valentin.malykh@phystech.edu