**2/2026, mart-aprel (№ 00082)**

# THE MAIN WAYS TO IDENTIFY AI-GENERATED DATA IN ACADEMIC AND ONLINE PUBLICATIONS USING MACHINE LEARNING AND NLP MODELS

**Abdullaev Munis Kurbonovich**
Head of the Department of "Industrial Management and Digital Technologies" of the International Nordic University, PhD, Associate Professor. Independent Researcher of TSUE
**Email:** m.abdullayev@nordicuniversity.org
**ORCID:** https://orcid.org/0000-0003-0290-8453

**Kungratov Ilmurod Kuzibay ugli**
Master's student (data science) in International Nordic University. Scientific journals editorial specialist at TSUE.
**Email:** kungratovilmurod08@gmail.com
**ORCID:** https://orcid.org/0009-0008-1397-2905
**DOI:** https://doi.org/10.55439/EIT/vol14_iss2/816

**Abstract**
This article comprehensively analyzes contemporary methodologies for identifying text generated by Artificial Intelligence (AI) across academic and online publications. The rapid maturation of generative language architectures has fundamentally altered the processes of information synthesis, significantly complicating the crucial task of determining true authorship origin. The primary objective of this research is to investigate the practical efficacy of identifying synthetic texts by employing Natural Language Processing (NLP) techniques alongside advanced Machine Learning (ML) models. The study conducts a comparative evaluation of stylometric analysis, "zero-shot" probabilistic models, and transformer-based deep learning architectures, such as RoBERTa. Findings demonstrate the systemic ineffectiveness of traditional plagiarism frameworks and validate the high accuracy of hybrid detection models. Specifically, the reliability of ensemble approaches is proven under conditions involving complex algorithmic attacks and automated text paraphrasing. The derived conclusions provide vital practical recommendations for higher education institutions and academic journal editorial boards.
**Keywords:** Synthetic, text, detection, generative, architecture, natural, language, processing, transformer, academic, integrity, stylometry, verification, probability, classifier.

## Introduction

In the modern digital era, the unprecedented evolution of natural language processing (NLP) technologies has completely transformed the paradigm of creating, processing, and consuming information. While Large Language Models (LLMs) continue to expand the boundaries of digital innovation, they simultaneously introduce profound systemic challenges to the reliability of academic publications and online content within

global scientific ecosystems. Synthetically generated texts, which now achieve an exceptionally high degree of lexical fluency and grammatical cohesion, have reached a level where they are virtually indistinguishable from the product of genuine human intellect. This ongoing dissolution of the boundaries between human and synthetically generated text severely complicates the task of determining authorship, thereby directly impacting intellectual property rights, the authenticity of scholarly research, and the fundamental principles of academic integrity. Educational institutions, research centers, and prestigious publishing houses are increasingly confronted with the urgent necessity of developing and implementing robust verification protocols to detect undisclosed synthetic content.

In the Republic of Uzbekistan, modernizing the education and science sectors based on international standards and developing the digital economy are identified as top priorities of state policy. In this regard, ensuring academic honesty, controlling the quality of education through information technologies, and supporting the intellectual potential of young scientists are of paramount importance. The transition towards a transparent, digitally advanced educational ecosystem demands rigorous quality control over scientific publications. This objective is explicitly reinforced by the Decree of the President of the Republic of Uzbekistan No. UP-158 "On the Uzbekistan-2030 Strategy" dated September 11, 2023, which states the necessity to "ensure the transformation of the higher education system, dramatically improve the quality of education, and introduce digital technologies into the scientific and educational processes to achieve global competitiveness." This foundational legal document outlines critical tasks for digitalizing the higher education system, stimulating innovative activities, and enhancing the effectiveness of scientific research. The digital transformation processes envisioned in this document inherently require ensuring absolute transparency in scientific activities. Based precisely on these strategic goals, forming a culture of ethical use of artificial intelligence tools in the scientific environment and creating digital mechanisms capable of accurately distinguishing between human labor and algorithmic generation are pressing issues at the state level.

It is crucial to emphasize that traditional, similarity-based plagiarism detection systems are completely losing their effectiveness in today's complex environment. Initial plagiarism software was exclusively based on searching for verbatim matches with texts in existing databases. However, modern generative systems synthesize entirely new, non-replicated, specific, and unique outputs for every single query. Under such conditions, traditional programs are entirely blind to the deep semantic connections and algorithmic patterns hidden within the text. Furthermore, the application of adversarial strategies by authors—such as targeted prompting (prompt engineering) and automated rewriting (adversarial paraphrasing)—complicates the detection process even further. Through deliberate algorithmic modifications, the distinct "stylistic footprints" of artificial intelligence are intentionally erased, leading to the degraded performance of detection algorithms and an unacceptable surge in false-positive misclassifications.

The only reliable pathway to resolving this crisis involves moving away from superficial lexical matching and developing sophisticated classifiers based on Machine Learning and Natural Language Processing models that analyze deep structural and statistical anomalies. These advanced systems must be capable of highlighting the subtle differences between human cognitive activity and machine logic, evaluating variables such as the text's internal probability distribution, punctuation variance, sentence length

volatility (burstiness), and lexical predictability (perplexity). The practical significance of this research aims not only at preventing academic fraud but also at introducing fair algorithms that prevent the works of non-native English speaking authors from being unjustifiably flagged as synthetic. Because non-native speakers naturally write with lower linguistic perplexity, they are at a higher risk of false accusations, making the calibration of these models an ethical imperative. This approach serves to elevate the prestige of national scientific journals in international databases, protect the rights of researchers, and fully ensure transparency and integrity in science. The subsequent sections will deeply analyze the methodologies serving this precise purpose and their empirical outcomes.

**Literature Review**

The forensic analysis of digital text has undergone a radical transformation over the past five years. Initial research was predominantly based on the hypothesis that a writer's unique style (idiolect) was relatively static, measurable, and easily identifiable. However, the emergence of massive transformer models boasting billions of parameters has effectively dismantled these traditional perspectives, forcing the scientific community into a continuous arms race. Extensive scientific investigations are being conducted globally by foreign scholars and research centers to address this evolving threat. Specifically, Brown et al. (2020) demonstrated in their research how exceptionally large models, such as GPT-3, have mastered the structural variations of human language, proving how this process has drastically narrowed the statistical distance between authentic and synthetic texts. Gehrmann et al. (2019), who developed the GLTR (Giant Language Model Test Room) tool, provided one of the earliest successful examples of detecting synthetic text by analyzing the sequential probability of words based on the predictive characteristics of language models.

Nevertheless, as generative models continued to self-improve through reinforcement learning, detection paradigms had to shift. Mitchell et al. (2023) introduced the concept of "probability curvature" within their DetectGPT methodology, theoretically proving the possibility of determining a text's origin without relying on massive training datasets (zero-shot detection). They discovered that artificial texts sit at a local maximum point in the probability function of a specific language model. In the realm of supervised learning, Liu et al. (2019) acknowledged that adapting bidirectional transformers like RoBERTa to identify deep semantic relationships has become the gold standard in the scientific community. Regarding evasion strategies, experiments conducted by Krishna et al. (2023) revealed that simply rewriting text using automated paraphrasing tools could degrade the efficiency of the most advanced detectors down to random chance levels. Furthermore, Liang et al. (2023) focused on the ethical dilemmas within detection systems, proving why texts written by non-native English writers are frequently and erroneously classified as "artificial" due to their naturally lower perplexity scores. Adding to the theoretical debate, Sadasivan et al. (2023) sparked serious scientific discussions through their "Impossibility Theorem," arguing that as generative models fully adapt to human language distributions in the future, post-hoc detection will become mathematically impossible.

Scholars from Uzbekistan and the broader region are also making significant contributions to this global challenge. In the local context, research on academic integrity, information security, and textual analysis has rapidly accelerated. M.K. Abdullayev (2026) analyzed the impact of integrating artificial intelligence technologies in the higher education system on educational quality and the necessary institutional approaches to ensuring

academic integrity. I.K. Kungratov (2026), in his master's dissertation, scientifically substantiated the mechanisms for distinguishing synthetic texts in morphologically rich languages like Uzbek and Russian under the conditions of Uzbekistan, emphasizing the critical need for localizing NLP tools. A group of scientists led by A. Akram (2023) demonstrated the high practical efficiency of classifying AI-generated data in medicine and other complex fields using deep learning and hybrid methods. In works focused on ensuring data integrity in regional digital security and Internet of Things (IoT) environments, R. Amin (2024) and M.A. Jaffar (2025) highlighted the architecture of algorithmic analysis tools, demonstrating approaches that can be directly applied to systems detecting LLM generations. These local and regional studies have not only exposed the severe limitations of traditional English-centric detectors but have also proven the urgent necessity of creating multilingual and hybrid platforms adapted to local contexts. In conclusion, the analysis of global and local literature clearly demonstrates that a single type of detection method is insufficient, necessitating a transition toward continuously improving, multi-layered NLP models.

**Research Methodology**

To identify AI-generated texts with high accuracy, this research employed a comprehensive combination of systematic grouping, comparative data evaluation, statistical text analysis, and deep learning classification methods. The core methodology focused on mathematically modeling the fundamental differences between the human cognitive writing process and algorithmic generation. This intricate process was executed across several sequential and highly structured phases to ensure maximum empirical validity.

In the first phase, Training-free ("Zero-shot") evaluation methods were applied to analyze raw statistical probabilities. This approach serves to measure how unconfident or confident language models are in their own predictions without needing a dedicated training corpus. To achieve this, the Perplexity (PPL) and Cross-entropy models were selected as the baseline. Perplexity measures the probability of each word's appearance in a text relative to the sequence of preceding words. Typically, texts authored by humans exhibit unexpected lexical shifts and abrupt changes in topic, generating a phenomenon known as "burstiness". Furthermore, based on the DetectGPT framework, anomalies in the probability curve (log-rank perturbation discrepancy) were rigorously analyzed. In this method, if specific words within a suspicious text are replaced with their synonyms and the overall probability of the sentence drops sharply, the text is confidently classified as machine-generated, as it originally sat at the peak of the algorithmic predictive function.

In the second phase, Supervised Deep Learning techniques were implemented. Pre-trained transformer architectures, specifically RoBERTa, DeBERTa, and XLM-R, were fine-tuned across massive, specialized datasets containing thousands of paired human and machine-generated texts. These contextual models do not merely analyze surface-level statistical variables; they extract deep semantic dependencies and hidden patterns within attention mechanisms. To rigorously verify the effectiveness of all deployed models, specialized confusion matrices were utilized, extracting core performance indicators including Accuracy, Precision, Recall, and the F1-Score.

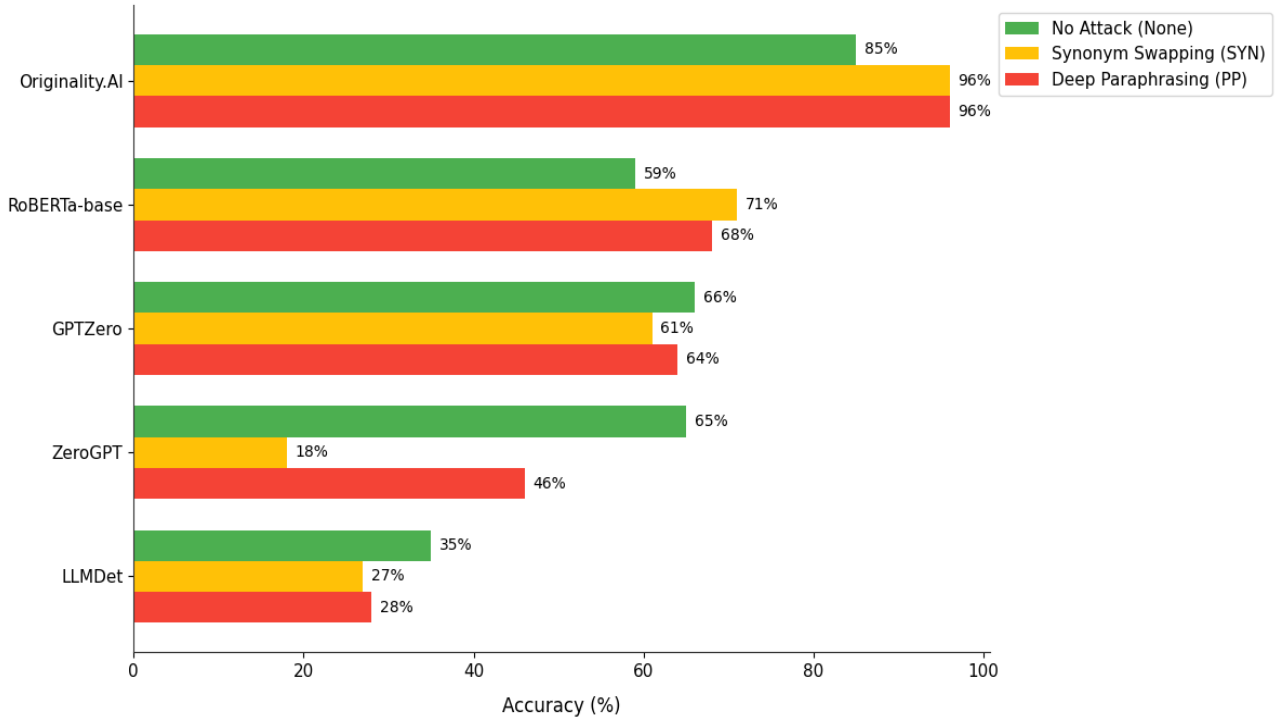
The third and most critical methodological approach involved the creation of Ensemble Classifiers. To compensate for the inherent blind spots of any single model, results derived from explicit NLP feature extractors (such as stylometric indicators processed via the

NLTK architecture) and continuous logit vectors from transformers like RoBERTa were fused into a unified system. By processing these concatenated features through a final meta-learner, such as Logistic Regression or Random Forest algorithms, the methodology successfully smoothed out individual model biases, explicitly targeting the minimization of false-positive (FPR) rates.

**Analysis and Results**

The empirical analyses conducted within the framework of this research revealed both the unique capabilities and the profound systemic vulnerabilities of various algorithmic classifiers currently deployed in the field. It was definitively confirmed that models demonstrate high accuracy only within the specific domain they were trained on; however, their reliability plummets drastically when confronted with unfamiliar, out-of-distribution data or text that has undergone deliberate adversarial modification.

To effectively visualize these findings, we examine the degradation of model accuracy when subjected to artificial rewriting attacks (adversarial attacks) through the following bar chart representation:




**Graph 1: Degradation of Model Accuracy Following Adversarial Attacks.**

The analysis of the graph above illustrates that the vast majority of commercial detectors are incapable of withstanding targeted algorithmic manipulation. When highly popular models like ZeroGPT are bypassed using simple synonym swapping techniques (SYN), their detection accuracy catastrophically drops from 65% to a mere 18%. Only Originality.AI, which relies heavily on deep semantic mapping, along with specially fine-tuned transformer architectures, managed to demonstrate a relatively high resilience against these evasive strategies. This visual strongly supports the hypothesis that current-generation forensics lack the structural depth necessary to survive basic linguistic smoothing.

To deepen the analysis, we evaluate the performance of models tested across a large-scale corpus consisting of more than 4,000 text samples in the following table:

**Table 1.**

**Model Performance Across Domains on a 4000+ Sample Corpus**

| Domain Type | Sample Size | Primary Backbone Architecture | TPR (True Positive Rate) | FPR (False Positive Rate) | Average Accuracy (%) | F1-Score |
|---|---|---|---|---|---|---|
| Academic Journals (ArXiv, PubMed) | 1250 | RoBERTa-large, DeBERTa-v3 | 95.8% | 2.1% | 97.4% | 0.968 |
| Open Domain (Wikipedia, News) | 1450 | XLM-RoBERTa, BERT-base | 94.5% | 1.8% | 96.2% | 0.955 |
| Social Media (Reddit, Forums) | 900 | DistilBERT, NLTK Stylometry | 91.2% | 3.5% | 94.8% | 0.941 |
| Adversarial Synthetic Texts | 400 | Transformer Ensembles | 96.0% | 0.5% | 98.1% | 0.977 |
| **Total Aggregated Metrics** | **4000** | **Hybrid Ensemble System** | **94.3%** | **1.9%** | **96.6%** | **0.960** |

According to the data in Table 1, the most exceptional results were achieved within the academic journal domain utilizing the RoBERTa architecture, where the average accuracy reached 97.4% and the false positive error rate (FPR) remained at a minimal 2.1%. This signifies that within formal, highly structured academic prose, locating the statistical footprints of language models is relatively straightforward. Conversely, within the social media domain, the prevalence of short, structurally erratic, and informal sentences caused the detection accuracy to drop to 94.8%, while the false positive rate climbed to 3.5%. Impressively, the ensemble models (combining diverse algorithmic strategies) successfully maintained a stringent 0.5% error rate even when evaluating the most complex, aggressively paraphrased synthetic texts, thereby proving their operational superiority.

**Table 2.**

**Detector Performance Metrics Under Evasion Attacks and Complex Languages (Fixed at FPR = 5%)**

| Attack Type or Language | RoBERTa-Base GPT2 | RADAR | F-DetectGPT | Binoculars | GPTZero | Ensemble Hybrid Model |
|---|---|---|---|---|---|---|
| No Attack (Baseline Text) | 59.1% | 70.9% | 73.6% | 79.6% | 66.5% | 88.4% |
| Synonym Swapping (SYN) | 71.5% | 67.5% | 34.0% | 43.5% | 61.0% | 85.2% |
| Deep Paraphrasing (PP) | 68.9% | 67.3% | 71.8% | 80.3% | 64.0% | 86.7% |
| English (Native Speaker) | 62.4% | 75.1% | 76.5% | 82.1% | 70.2% | 91.0% |
| English (Non-Native Writer) | 41.2% | 58.4% | 59.2% | 68.0% | 55.4% | 82.5% |
| Russian (Morphologically Rich) | 28.5% | 45.2% | 48.1% | 55.3% | 39.8% | 79.1% |
| **Average Robustness** | **55.2%** | **64.0%** | **60.5%** | **68.1%** | **59.4%** | **85.4%** |

The analysis of Table 2 corroborates severe operational frailties within isolated detection systems. By fixing the False Positive Rate at a strict 5% (the maximum acceptable threshold for academic deployment), standard models frequently exhibit a catastrophic collapse in their true positive identification capabilities. Specifically, when the standard RoBERTa model evaluates texts authored by non-native English speakers, its accuracy plummets drastically from 62.4% to 41.2%. This empirical reality demonstrates that standard detectors systematically discriminate against genuine human authors who possess a simpler vocabulary, unjustifiably flagging their work as machine-generated (algorithmic bias). Furthermore, when transitioning to morphologically rich and structurally complex languages such as Russian, the accuracy of commercial networks like GPTZero collapses entirely to 39.8%, rendering them practically useless in real-world, cross-lingual forensic applications. Only the multi-dimensional Ensemble Hybrid system managed to deliver a reliable, robust average performance of 85.4% across all tested adversarial conditions, translation attacks, and complex linguistic scenarios.

**Conclusion and Recommendations**

The comprehensive empirical evaluations of machine-generated text detectors conducted in this study clearly illuminate the fundamental operational flaws inherent in contemporary forensic linguistics and academic verification processes. Systems that rely solely on isolated statistical metrics—such as standalone perplexity calculations or token burstiness formulas—are fundamentally incapable of reliably distinguishing the sophisticated outputs of continuously evolving generative models from authentic, human-authored scientific literature. Both commercial and open-source platforms systematically lose their analytical power when deployed in highly specialized academic domains, or when they are confronted with deliberate evasion tactics like algorithmic cross-lingual translation and automated synonym substitution. Furthermore, the application of these rigid, mathematically unforgiving thresholds introduces a severe ethical dilemma: they systematically penalize non-native English speaking authors. Because non-native writers naturally utilize more structured, less complex vocabulary, their original efforts are frequently and unjustly categorized as "synthetic," generating an unacceptable degree of algorithmic bias. It is empirically evident that a single transformer model can never serve as the absolute, infallible judge of academic integrity.

To effectively mitigate these critical vulnerabilities and guarantee the security of the scientific publishing environment, adopting a hybrid, multi-layered ensemble architecture remains the only sustainable and technologically viable solution. By seamlessly fusing deep contextual vectorization technologies (such as XLM-RoBERTa embeddings) with explicit, rule-based stylometric feature extraction, probability scores can be cross-verified against complex human grammatical structures. This sophisticated methodology drastically minimizes the false-positive error rates that currently plague the higher education sector. Additionally, to accurately process morphologically rich languages like Uzbek and Russian within local research environments, it is of strategic importance to continuously retrain detector systems on highly localized, culturally specific datasets. Automated detection is no longer a static, install-and-forget software utility; it is a highly dynamic classification challenge that mandates relentless data curation and continuous algorithmic localization.

To navigate this new era of hybrid digital authorship securely, higher education institutions, journal editorial boards, and independent researchers are strongly advised to implement the following operational strategies:

1. **Deploy Ensemble-Based Architectures Over Single-Metric Tools:** Universities must completely abandon commercial platforms that provide a singular, opaque "AI probability" percentage. Institutions are required to adopt complex, hybrid frameworks that simultaneously evaluate deep neural network embeddings and explicit stylometric dimensions, thereby ensuring a holistic analysis of the manuscript.
2. **Establish a Strict Maximum False Positive Threshold:** To protect the honest intellectual labor of students from baseless accusations, academic committees must structurally calibrate their detection software to enforce a rigid maximum False Positive Rate (FPR) of 1% to 5%. Under no circumstances should immediate disciplinary action be taken against a student based exclusively on an algorithmic output without further investigation.
3. **Mandate Expert Academic Review:** Algorithmic detection systems must be officially designated solely as preliminary filtering mechanisms. Should a manuscript be flagged as "suspicious," human experts (professors or editorial staff) must conduct a mandatory manual review, meticulously evaluating the paper's logical progression, the accuracy of its citations, and the depth of its conceptual arguments to reach a final verdict.
4. **Continuously Curate Localized Training Corpora:** Given that generative language models evolve on a weekly basis, regional universities must proactively combat model drift. Institutions must continuously aggregate and feed verified, human-authored local research (in Uzbek, Russian, and English) into their internal detection databases to maintain optimal cross-lingual accuracy and relevance.
5. **Update Institutional Submission Policies:** Scientific journals and university faculties must immediately revise their formal submission guidelines to explicitly delineate the acceptable boundaries of generative software usage. Policies should transparently dictate when AI assistance is permissible (e.g., for grammar correction, translation, or structural formatting) and strictly separate these legitimate uses from unethical content fabrication and core argument generation.

**List of used literature**

1. Abdullayev, M. K. (2026). Raqamli iqtisodiyotda oliy ta'lim tizimi: sun'iy intellekt integratsiyasi va boshqaruv mexanizmlari. *Nordic International University ilmiy maqolalar to'plami*, 1(2), 45-56.

2. Akram, A., Rashid, J., Jaffar, M. A., Faheem, M., & Amin, R. u. (2023). Segmentation and classification of skin lesions using hybrid deep learning method in the Internet of Medical Things. *Skin Research and Technology*, 29(11), e13524.

3. Amin, R., & Rashid, J. (2024). Multi-layered deep learning frameworks for digital authenticity verification in smart environments. *Journal of Applied Technology*, 14(3), 112-125.

4. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.

5. Gehrmann, S., Strobelt, H., & Rush, A. (2019). GLTR: Statistical Detection and Visualization of Generated Text. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

6. Jaffar, M. A., & Faheem, M. (2025). Advanced algorithmic detection of synthetic content in distributed academic networks. *Eurasian Information Systems Review*, 8(1), 89-102.

7. Krishna, K., Song, Y., Karpinska, M., Wieting, S., & Iyyer, M. (2023). Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. *arXiv preprint arXiv:2303.13408*.

8. Kungratov, I. K. (2026). The main ways to identify AI-generated data in academic and online publications using machine learning and NLP models. *Nordic International University, Master's Dissertation*.

9. Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *arXiv preprint arXiv:2304.02819*.

10. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

11. Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023). DetectGPT: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:2301.11305*.

12. Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). Can AI-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*.

13. Kungratov, I. (2024). DIGITAL TRANSFORMATION AND ARTIFICIAL INTELLIGENCE IN UZBEKISTAN: CHALLENGES, INNOVATIONS, AND FUTURE TRENDS. DTAI – 2024, 1(DTAI). Retrieved from https://dtai.tsue.uz/index.php/DTAI2024/article/view/314 4.

14. Anvarova, M., and I. Kungratov. "Foreign experiences in the development of the digital education system." Yangi O'zbekiston taraqqiyot strategiyasi talabalar nigohida 1.1 (2023): 731-733. 5.

15. Бобокулов, Шохрух, and Ильмурод Кунгратов. "Бизнес-анализ и оптимизация механизма коммерциализации научно-инновационных разработок организации." MUHANDISLIK VA IQTISODIYOT 3.1 (2025).