УДК 004.85
DOI https://doi.org/10.32782/IT/2024-4-2

Катерина АНТІПОВА
доктор філософії, старший викладач кафедри інженерії програмного забезпечення, Чорноморський національний університет імені Петра Могили, вул. 68 Десантників, 10, м. Миколаїв, Україна, 54000
***ORCID:** 0000-0002-9012-5290*
***Scopus Author ID:** 57212609599*

Ігор КАНДИБА
доктор філософії, старший викладач кафедри інженерії програмного забезпечення, Чорноморський національний університет імені Петра Могили, вул. 68 Десантників, 10, м. Миколаїв, Україна, 54000
***ORCID:** 0000-0002-8589-4028*
***Scopus Author ID:** 57212577217*

Світлана БОРОВЛЬОВА
старший викладач кафедри інженерії програмного забезпечення, Чорноморський національний університет імені Петра Могили, вул. 68 Десантників, 10, м. Миколаїв, Україна, 54000
***ORCID:** 0000-0003-1994-0556*

Віктор РАЛЕНКО
викладач кафедри інженерії програмного забезпечення, Чорноморський національний університет імені Петра Могили, вул. 68 Десантників, 10, м. Миколаїв, Україна, 54000
***ORCID:** 0009-0009-4161-8468*

Бібліографічний опис статті: Антіпова, К., Кандиба, І., Боровльова, С., Раленко, В. (2024). Виявлення плагіату в тексті, згенерованого великими мовними моделями. *Information Technology: Computer Science, Software Engineering and Cyber Security,* 4, 9–15, doi: https://doi.org/10.32782/IT/2024-4-2

ВИЯВЛЕННЯ ПЛАГІАТУ В ТЕКСТІ, ЗГЕНЕРОВАНОГО ВЕЛИКИМИ МОВНИМИ МОДЕЛЯМИ

Для виявлення тексту, згенерованого за допомогою великих мовних моделей, існують різні підходи, на основі яких розроблені такі алгоритми, як DetectGPT, RADAR, Ghostbuster, GPT-Sentinel та інші для виявлення контенту, згенерованого штучним інтелектом. Хоча автоматизована перевірка може виявити певний плагіат, дослідження показують, що програмне забезпечення для пошуку текстів не тільки не знаходить весь плагіат, але й позначає оригінальний контент як плагіат, надаючи таким чином хибнопозитивні результати. Найсучасніші детектори згенерованого тексту демонструють значне погіршення продуктивності, коли стикаються з текстами, створеними людьми, які не є носіями англійської мови.

***Метою дослідження є** підвищення точності і надійності виявлення тексту, створеного штучним інтелектом, особливо в освітньому середовищі, де плагіат та академічна недоброчесність стають усе більш актуальними через використання генеративних мовних моделей.*

***Методологія дослідження** базується на загальнонаукових методах аналізу та синтезу, експериментальному тестуванні та кількісному аналізі ефективності мовної моделі, призначеної для перевірки тексту на наявність плагіату.*

***Наукова новизна дослідження** полягає в адаптації сучасних методів виявлення плагіату для надійної класифікації текстів, створених штучним інтелектом, у контексті української мови. Для цього створено новий датасет на основі перефразованих текстових фрагментів, згенерованих ChatGPT, та налаштовано модель для класифікації тексту. Ефективність моделі оцінена за допомогою трьох різних оціночних метрик: показника F1, частоти хибно позитивних спрацювань і частоти хибно негативних спрацювань.*

***Результати дослідження показують**, що налаштована модель ефективно виявляє відмінності між двома типами тексту, надають певне представлення про сильні та слабкі сторони моделі і демонструють її потенціал для застосування для практичних задач. Подальше дослідження полягає у зборі даних з іншим контекстом, щоб оцінити точність налаштованої моделі для різних задач обробки природної мови.*

***Ключові слова:** ChatGPT, академічна недоброчесність, велика мовна модель, виявлення плагіату, генерація тексту, налаштування моделі.*

Kateryna ANTIPOVA
PhD, Senior Lecturer at the Department of Software Engineering, Petro Mohyla Black Sea National University, 10, 68 Desantnykiv Str., Mykolaiv, Ukraine, 54000
***ORCID:** 0000-0002-9012-5290*
***Scopus Author ID:** 57212609599*

Ihor KANDYBA
PhD, Senior Lecturer at the Department of Software Engineering, Petro Mohyla Black Sea National University, 10, 68 Desantnykiv Str., Mykolaiv, Ukraine, 54000
***ORCID:** 0000-0002-8589-4028*
***Scopus Author ID:** 57212577217*

Svitlana BOROVLOVA
Senior Lecturer at the Department of Software Engineering, Petro Mohyla Black Sea National University, 10, 68 Desantnykiv Str., Mykolaiv, Ukraine, 54000
***ORCID:** 0000-0003-1994-0556*

Viktor RALENKO
Lecturer at the Department of Software Engineering, Petro Mohyla Black Sea National University, 10, 68 Desantnykiv St., Mykolaiv, Ukraine, 54000
***ORCID:** 0009-0009-4161-8468*

To cite this article: Antipova, K., Kandyba, I., Borovlova, S., & Ralenko, V. (2024). Vyiavlennia plahiatu v teksti, zghenerovanoho velykymy movnymy modeliamy [Plagiarism detection in texts generated by large language models]. *Information Technology: Computer Science, Software Engineering and Cyber Security,* 4, 9–15, doi: https://doi.org/10.32782/IT/2024-4-2

PLAGIARISM DETECTION IN TEXTS GENERATED BY LARGE LANGUAGE MODELS

Various algorithms have been developed to detect text generated by large language models, and instruments such as DetectGPT, RADAR, Ghostbuster, GPT-Sentinel, etc. are based on these algorithms. While automated systems can detect some plagiarism, studies show that such software not only fails to find all plagiarism, but also marks original content as plagiarized, thus providing false positive results. State-of-the-art AI-text detectors demonstrate a significant performance degradation when faced with texts created by non-native English speakers.

***The purpose of this study** is to improve the accuracy and reliability of detecting AI-generated text, especially in the educational environment, where plagiarism and academic misconduct are becoming increasingly relevant due to the use of LLMs.*

***The research methodology** is based on general scientific methods of analysis and synthesis, experimental testing, and quantitative analysis of the effectiveness of a language model fine-tuned for plagiarism detection.*

***The scientific originality of the study** lies in the adaptation of modern methods of plagiarism detection for reliable classification of AI-texts in the context of the Ukrainian language. To do this, a new dataset was created based on paraphrased text fragments generated by ChatGPT, and the mT5 model was fine-tuned on this dataset for a classification task. The model's performance is evaluated using three different evaluation metrics: F1 score, false positive rate, and false negative rate.*

***The results of the study** show that the fine-tuned model effectively detects the differences between the two types of text. The results also provide some insight into the strengths and weaknesses of the model, and demonstrate its potential for application in practical tasks. Further research aims to collect data from different contexts to evaluate the accuracy of the fine-tuned model for different natural language processing tasks.*

***Key words:** Academic misconduct, ChatGPT, fine-tuning, LLM, plagiarism detection, text generation.*

Актуальність проблеми. Впровадження ChatGPT, революційного інструменту на базі великої мовної моделі (ВММ), значно змінило різні галузі та сфери, включаючи й академічну. Можливості розвинутих ВММ вплинули на академічний світ різними способами. Наприклад, здобувачі вищої освіти використовують ChatGPT для виконання домашніх завдань і складання іспитів. Це викликало занепокоєння стосовно поточних систем оцінювання, які використовуються у закладах вищої освіти. Викладачі та університети намагаються виявити шахрайські дії здобувачів, і плагіат є однією з головних проблем. Раніше плагіат здебільшого полягав

у представленні рефератів та есе, які містили абзаци з інших джерел без посилань на них, але з появою ВММ здобувачі тепер можуть використовувати штучний інтелект (ШІ) для створення тексту і виконання своїх завдань. Акт використання здобувачами тексту, згенерованого ВММ, і ствердження, що це їхня власна робота, називається ШІ-плагіатом. Залежність здобувачів від інструментів генерації тексту призводить до втрати креативності та здатності до навчання.

Мовні моделі – це засновані на глибокому машинному навчанні моделі, призначені для різних завдань з обробки природної мови. Такі ВММ, як ChatGPT, можуть перефразовувати текст і створювати такий текст, що майже не відрізняється від тексту, написаного людиною. Ці моделі можуть впоратися як з простими завданнями, такими як створення есе на задану тему, так і зі складними, наприклад, написання наукової роботи зі складної проблеми. Поява таких генеративних інструментів ШІ та їхня здатність генерувати схожий на людський текст становить значну загрозу для академічної доброчесності.

Завдання виявлення того, чи був певний текст згенерований алгоритмами на основі ШІ, називається виявленням контенту, згенерованого штучним інтелектом (КЗШІ). Починаючи з моделі ELMO з 94 мільйонами навчальних параметрів у 2019 році до сучасної моделі GPT-4 з 1.76 трильйонами параметрів, еволюція ВММ та їхні можливості постійно зростають. З прогнозованим швидким розвитком високопродуктивних ВММ якість вихідних текстів зростає, що ускладнює їх виявлення.

Аналіз останніх досліджень і публікацій. Для виявлення тексту, згенерованого за допомогою ВММ, були розроблені різні підходи, такі як навчання класифікаторів, додавання водяних знаків та «безпрограшні» підходи (*zero-shot approaches*). На основі цих підходів розробляються такі алгоритми, як DetectGPT (Mitchell, 2023), RADAR (Hu, 2023), Ghostbuster (Verma, 2023), GPT-Sentinel (Chen, 2023) та інші для виявлення контенту, згенерованого ШІ. OpenAI представив свій інструмент виявлення КЗШІ через два місяці після випуску ChatGPT. Однак OpenAI стверджує, що детектор не є повністю надійним. Аналогічно, кілька інструментів і програм для виявлення КЗШІ, таких як CopyLeaks, Turnitin, GPTZero і Crossplag, були випущені для загального користування для виявлення контенту, згенерованого ШІ. Підходи до розпізнавання машинного тексту можна розділити на 4 категорії (Chen, 2023).

Традиційний статистичний підхід шляхом аналізу статистичних аномалій у текстовій вибірці. Деякі статистичні показники, такі як ентропія, використовуються як порогові значення для розпізнавання ШІ-тексту. Типовим прикладом є DetectGPT (Mitchell, 2023). DetectGPT перетворює вхідний текст за допомогою мовної моделі, яка заповнює маску, наприклад, T5. Потім виконується виявлення ШІ-тексту шляхом порівняння ймовірностей для тексту та його заповнених варіантів.

Підхід з неконтрольованим навчанням шляхом класифікації ВММ з нуля. Виявлення ШІ-тексту формулюється як задача бінарної класифікації, а класифікатор навчається для цільової мовної моделі (Rodriguez, 2022). Наприклад, OpenAI навчає свій класифікатор ШІ-тексту за моделлю на основі RoBERTa.

Підхід з контрольованим навчанням шляхом налаштування (*fine-tuning*) мовної моделі з додаванням або без додавання модуля класифікації.

Постфактумні методи нанесення водяних знаків, такі як методи на основі глибокого навчання (Dai, 2022), можуть бути застосовані до ВММ. У (Kirchenbauer, 2023) запропоновано м'яку схему водяного маркування, яка вбудовує водяний знак у кожне слово згенерованого речення, розділяючи словниковий запас на різні списки і диференційовано вибираючи наступний маркер. Однак, як показано в (Sadasivan, 2023), ефективність багатьох існуючих детекторів ШІ-контенту значно знижується через перефразування тексту.

Метою дослідження є підвищення точності і надійності виявлення тексту, створеного ШІ, особливо в освітньому середовищі, де плагіат та академічна недоброчесність стають усе більш актуальними через використання генеративних мовних моделей. Дослідження спрямоване на адаптацію сучасних методів виявлення плагіату для надійної класифікації текстів, створених ШІ, у контексті української мови. Для цього створено новий датасет на основі перефразованих текстових фрагментів, згенерованих ChatGPT, та налаштовано модель mT5 для класифікації тексту.

Виклад основного матеріалу. Академічна недоброчесність – це дії, які порушують оригінальність академічної роботи, такі як гострайтинг (*ghostwriting*), плагіат, фабрикація даних, генерація із використанням ШІ. Серед цих порушень найпоширенішою і найновішою формою неправомірної поведінки є генерація текстів. Дослідження в першу чергу зосереджується на

цьому аспекті. Найновіші інструменти генеративного ШІ, такі як ChatGPT, GPT-4 Vision, SORA від OpenAI, Gemini від Google та Perplexity від Perplexity AI, здатні генерувати різні типи контенту, включаючи текст, зображення, відео та код на різних мовах програмування.

Типологія плагіату може відрізнятися залежно від типу даних або рівня перетворення тексту. Фолтнек та ін. (Foltynek, 2019) представили різні типології, визначені в кількох наукових роботах, і запропонували нову типологію плагіату за рівнем перетворення: плагіат зі збереженням символів, зі збереженням синтаксису, зі збереженням семантики, зі збереженням ідеї та гострайтинг. Тип плагіату може також відрізнятися залежно від типу даних, наприклад, плагіат коду та плагіат тексту. Виявлення плагіату може бути зовнішнім або внутрішнім. Якщо плагіат виявляється лише за допомогою самого тексту, то це внутрішнє виявлення. Тоді якщо плагіат виявляється в порівнянні з іншим текстом, це називається зовнішнім виявленням. Виявлення плагіату також можна класифікувати відповідно до підходу: векторного, синтаксичного, семантичного, нечіткого, структурного та стиліметричного (Pudasaini, 2024).

Типовий алгоритм виявлення плагіату включає в себе інженерію ознак, класифікаційні моделі або метрики схожості текстів. Загальні ознаки, що використовуються в більшості алгоритмів виявлення плагіату, – це частота символів, середня довжина слова, середня довжина речення, частота N-грам у слові, частина мови, синоніми та гіперніми. Плагіат в основному оцінюється на основі текстової схожості з іншими еталонними текстами. Для обчислення такої схожості найчастіше використовують відстань Хаммінга, відстань Левенштейна та відстань найдовшої спільної підпослідовності, тоді як коефіцієнт Жаккара, косинусний коефіцієнт, відстань Манхеттена, евклідова відстань, коефіцієнт відповідності та коефіцієнт Соренсена були найпоширенішими метриками векторної схожості (Pudasaini, 2024).

Хоча автоматизована перевірка може виявити певний рівень плагіату, дослідження показують, що програмне забезпечення для пошуку текстів не тільки не знаходить весь плагіат, але й позначає оригінальний контент як плагіат, надаючи таким чином хибнопозитивні результати. Інструменти виявлення текстів, створених ШІ, дають збої, вони не є ані точними, ані надійними (усі мають точність нижче 80%). Загалом було виявлено, що вони діагностують документи, написані людиною, як створені штучним інтелектом (хибнопозитивні результати) і часто діагностують тексти, створені ШІ, як написані людиною (хибнонегативні результати). Інструменти виявлення здебільшого схильні класифікувати результати як написані людиною, а не виявляти вміст, створений ШІ. Загалом, приблизно 20% текстів, згенерованих штучним інтелектом, можуть бути помилково приписані людині.

Потреба в розробці нових алгоритмів, здатних точно відрізняти машинний текст від написаного людиною, стала як ніколи актуальною. Сучасний детектор контенту, згенерованого ВММ, повинен володіти такими ключовими характеристиками (Jawahar, 2020):

– ***точність*** означає, що модель повинна бути здатна розрізняти текст, згенерований ВММ, і текст, написаний людиною, досягаючи при цьому відповідного компромісу між точністю і швидкістю відкликання;

– ***ефективність*** використання даних означає, що детектор повинен оперувати якомога меншою кількістю прикладів з мовної моделі;

– ***узагальненість*** означає, що детектор повинен працювати послідовно, незалежно від будь-яких змін в архітектурі моделі, довжини запиту або набору навчальних даних;

– ***інтерпретованість*** означає, що детектор повинен надавати чіткі пояснення причин, що лежать в основі його рішень.

Нещодавнє дослідження (Hu, 2023) показало, що найсучасніші детектори ШІ-тексту демонструють значне погіршення продуктивності, коли стикаються з текстами, створеними людьми, які не є носіями англійської мови. Отже, незважаючи на багатообіцяючі результати, сучасні моделі мають певні обмеження. Одним з них є те, що більшість моделей навчаються лише на англомовному корпусі текстів. Як наслідок, ефективність розпізнавання текстів іншими мовами, зокрема слов'янськими, не є оптимальною. Для усунення цього обмеження може бути корисним налаштування моделей на текстах українською.

Щоб вирішити цю проблему, ми обрали підхід з контрольованим навчанням для розрізнення тексту, написаного людиною, і тексту, згенерованого ШІ. Спочатку ми зібрали дані з ChatGPT, а потім налаштували мовну модель на цьому наборі даних для класифікації. ChatGPT – це мовна модель, заснована на архітектурі GPT-4. Наскільки нам відомо, наразі не існує загальнодоступного набору даних, який би систематично збирав тексти, що генеруються ChatGPT. Тому ми поставили собі завдання створити власний набір даних згенерованих ШІ.

Набір даних GPTText складається з перефразованих фрагментів тексту, які були згенеровані мовною моделлю з використанням датасету UberText 2.0 (Chaplynskyi, 2023) як джерела. Датасет UberText був взятий із загальнодоступного ресурсу, який містить кілька наборів даних сучасною українською: новини, художня література, тексти на соціальну та судову тематику, статті з Вікіпедії. Набір даних містить 26 819 текстових фрагментів, кожен з яких відповідає фрагменту тексту з датасету UberText. Фрагменти, довжина яких перевищувала 2 000 слів, були відфільтровані. В процедурі перефразування був використаний API OpenAI на моделі gpt-4.

Ми налаштували модель mT5 (Xue, 2020) для задач класифікації, від послідовності до послідовності (*seq-to-seq*). Модель складається з двох основних компонентів: блоку кодера та блоку декодера, кожен з яких повторюється шість разів. Блок кодера обробляє вхідні токени, використовуючи механізм самоуваги (*self-attention*), після чого застосовується багатошаровий персептрон з прямим поширенням. Блок декодера аналогічно застосовує маскована багатоголову увагу (*multi-head attention*) та шари з прямим поширенням до закодованих представлень, що дозволяє йому генерувати виходи токен за токеном, передбачаючи ймовірність наступного слова. Під час навчання вхідна послідовність складається з текстової вибірки з GPTText, а вихідна представляє результат класифікації у вигляді «pos </s>» або «neg </s>», де «</s>» є маркером кінця послідовності. Остаточний результат – це ймовірнісний розподіл для наступного слова, що використовується для розрізнення послідовностей, згенерованих людиною або ChatGPT.

Ефективність моделі mT5 оцінена за допомогою трьох різних оціночних метрик: показника F1, частоти хибно позитивних спрацювань (FPR – *false positive rate*) і частоти хибно негативних спрацювань (FNR – *false negative rate*). Тут «позитивний» означає, що вхідний текст згенерований ChatGPT, тоді як «негативний» означає, що дані написані людиною. Враховуючи кількість істинно позитивних (*TP*), істинно негативних (*TN*), хибно позитивних (*FP*) та хибно негативних (*FN*) результатів, метрики обчислюються наступним чином:

$$\mathrm{F1\,Score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad \mathrm{TPR} = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN} \quad TNR = \frac{TN}{TN + FP} \quad FNR = \frac{FN}{FN + TP}$$

За допомогою показника F1 вимірюють ступінь подібності між маркованою та передбачуваною відповіддю, отриманою від моделі. Показники F1, FPR і FNR для моделі mT5 розраховані на датасеті GPTText. Результати оцінювання наведені в таблиці 1. Всі дані подані у відсотках.

Таблиця 1

Показники точності моделей

Модель	Точність	TPR	TNR	FPR	FNR
mT5	95.98	97.3	96.22	3.78	2.7
GPT-2	41.5	22.38	95.87	4.13	77.62
GPT Zero	56.4	35.71	81.64	18.36	64.29

Результати дослідження показують, що налаштована модель ефективно виявляє відмінності між двома типами тексту, надає певне представлення про сильні та слабкі сторони моделі і демонструє її потенціал для застосування для практичних задач.

Висновки. В цій роботі за допомогою мовної моделі визначені відмінності між текстом, згенерованим за допомогою ChatGPT, і текстом, написаним людиною. Для цього був зібраний датасет, який складається з перефразованого контенту, згенерованого за допомогою ChatGPT. Після чого була налаштована модель mT5 для класифікації тексту. Ця модель досягла чудових результатів, з точністю понад 95% на тестовому наборі даних, оцінених за допомогою різних метрик. Такі результати дають важливу інформацію про ефективне використання мовних моделей для розпізнавання згенерованого тексту.

Однак, модель, навчена для задачі класифікації на такому наборі даних, як GPTText, може не дуже добре справлятися з іншими завданнями обробки природної мови, для яких ChatGPT широко використовується, наприклад, з відповідями на запитання. У майбутньому планується зібрати датасети з іншим контекстом, щоб оцінити точність налаштованої моделі mT5 для різних задач.

ЛІТЕРАТУРА:

1. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature / E. Mitchell et al. *Proceedings of Machine Learning Research :* Proceedings of the 40th International Conference on Machine Learning, 3 July 2023. Honolulu, Hawaii, USA, 2023. P. 24950-24962. DOI: 10.48550/arXiv.2301.11305

2. RADAR: Robust AI-Text Detection via Adversarial Learning / X. Hu et al. *Advances in Neural Information Processing Systems*, 10-16 December 2023. Vol. 36. New Orleans, USA, 2023. P. 15077–15095. DOI: 10.48550/arXiv.2307.03838

3. Ghostbuster: Detecting Text Ghostwritten by Large Language Models / V. Verma et al. *North American Chapter of the Association for Computational Linguistics*. USA, 2023. 16 p. (arXiv preprint). DOI: 10.48550/arXiv.2305.15047

4. GPT-Sentinel: Distinguishing Human and ChatGPT Generated Content / Y. Chen et al. 2023. 18 p. (arXiv preprint). DOI: 10.48550/arXiv.2305.07969

5. Cai S., Cui W. Evade ChatGPT Detectors via A Single Space. 2023. 12 p. (arXiv preprint). DOI: 10.48550/arXiv.2307.02599

6. Cross-domain detection of gpt-2-generated technical text / J. Rodriguez et al. *Human Language Technologies*: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics, July 2022. Seattle, USA, 2022. P. 1213–1233. URL: https://aclanthology.org/2022.naacl-main.88 (дата звернення: 07.09.2024)

7. DeepHider: A Multi-module and Invisibility Watermarking Scheme for Language Model / L. Dai et al. 2022. 16 p. (arXiv preprint). DOI: 10.48550/arXiv.2208.04676

8. A Watermark for Large Language Models / J. Kirchenbauer et al. *Proceedings of Machine Learning Research :* Proceedings of the 40th International Conference on Machine Learning, 3 July 2023. Honolulu, Hawaii, USA, 2023. 26 p. DOI: 10.48550/arXiv.2301.10226

9. Can AI-Generated Text be Reliably Detected? / V.S. Sadasivan et al. 2023. 34 p. (arXiv preprint). DOI: 10.48550/arXiv.2303.11156

10. Large Language Models can be Guided to Evade AI-Generated Text Detection / N. Lu et al. 2023. 29 p. (arXiv preprint). DOI:10.48550/arXiv.2305.10847

11. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense / K. Krishna et al. *Advances in Neural Information Processing Systems*, 10-16 December 2023. Vol. 36. New Orleans, USA, 2023. P. 27469-27500. DOI: 10.48550/arXiv.2303.13408

12. Authorship Obfuscation in Multilingual Machine-Generated Text Detection / D. Macko et al. 2024. 21 p. (arXiv preprint). DOI: 10.48550/arXiv.2401.07867

13. Foltynek T., Meuschke N., Gipp B. Academic plagiarism detection: a systematic literature review. *Association for Computing Machinery Computing Surveys (CSUR)*. Vol. 52(6). USA, 2019. p. 1–42. DOI: 10.1145/3345317

14. Survey on Plagiarism Detection in Large Language Models: The Impact of ChatGPT and Gemini on Academic Integrity / S. Pudasaini et al. 2024. 23 p. (arXiv preprint). URL: https://arxiv.org/pdf/2407.13105 (дата звернення: 10.10.2024)

15. Jawahar G., Abdul-Mageed M., Lakshmanan L. V. Automatic detection of machine generated text: A critical survey. *Proceedings of the 28th International Conference on Computational Linguistics,* December 2020. Barcelona, Spain, 2020. P. 2296–2309. URL: https://aclanthology.org/2020.coling-main.208 (дата звернення: 10.10.2024)

16. Chaplynskyi D. Introducing UberText 2.0: A Corpus of Modern Ukrainian at Scale. *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, May 2023. Dubrovnik, Croatia, 2023. P. 1–10. URL: https://aclanthology.org/2023.unlp-1.1 (дата звернення: 21.09.2024)

17. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer / L. Xue et al. *Human Language Technologies :* Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics, June 2021. USA, 2021. P. 483-498. DOI: 10.18653/v1/2021.naacl-main.41

18. GPT-2 Output Detector Demo. URL: https://openai-openai-detector.hf.space/ (дата звернення: 27.08.2024)

19. AI Detector – the Original AI Checker for ChatGPT & More. URL: https://gptzero.me/ (дата звернення: 27.08.2024)

REFERENCES:

1. Mitchell, E., Lee, Y., Khazatsky, A., Manning, C.D., & Finn, C. (2023). DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. *International Conference on Machine Learning*. https://doi.org/10.48550/arXiv.2301.11305

2. Hu, X., Chen, P., & Ho, T. (2023). RADAR: Robust AI-Text Detection via Adversarial Learning. https://doi.org/10.48550/arXiv.2307.03838

3. Verma, V. K., Fleisig, E., Tomlin, N., & Klein, D. (2023). Ghostbuster: Detecting Text Ghostwritten by Large Language Models. *North American Chapter of the Association for Computational Linguistics*. https://doi.org/10.48550/arXiv.2305.15047

4. Chen, Y., Kang, H., Zhai, V., Li, L., Singh, R., & Ramakrishnan, B. (2023). GPT-Sentinel: Distinguishing Human and ChatGPT Generated Content. https://doi.org/10.48550/arXiv.2305.07969

5. Cai, S., & Cui, W. (2023). Evade ChatGPT Detectors via A Single Space. https://doi.org/10.48550/arXiv.2307.02599

6. Rodriguez, J., Hay, T., Gros, D., Shamsi, Z., & Srinivasan, R. (2022). Cross-domain detection of gpt-2-generated technical text. In *NAACL*, pp. 1213–1233. https://aclanthology.org/2022.naacl-main.88

7. Dai, L., Mao, J., Fan, X., & Zhou, X. (2022). DeepHider: A Multi-module and Invisibility Watermarking Scheme for Language Model. https://doi.org/10.48550/arXiv.2208.04676

8. Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (2023). A Watermark for Large Language Models. *International Conference on Machine Learning*. https://doi.org/10.48550/arXiv.2301.10226

9. Sadasivan, V.S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). Can AI-Generated Text be Reliably Detected? https://doi.org/10.48550/arXiv.2303.11156

10. Lu, N., Liu, S., He, R., & Tang, K. (2023). Large Language Models can be Guided to Evade AI-Generated Text Detection. *Trans. Mach. Learn. Res*. https://doi.org/10.48550/arXiv.2305.10847

11. Krishna, K., Song, Y., Karpinska, M., Wieting, J., & Iyyer, M. (2023). Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. https://doi.org/10.48550/arXiv.2303.13408

12. Macko, D., Móro, R., Uchendu, A., Srba, I., Lucas, J.S., Yamashita, M., Tripto, N.I., Lee, D., Simko, J., & Bieliková, M. (2024). Authorship Obfuscation in Multilingual Machine-Generated Text Detection. https://doi.org/10.48550/arXiv.2401.07867

13. Foltynek, T., Meuschke, N., & Gipp, B. (2019). Academic plagiarism detection: a systematic literature review. *ACM Computing Surveys (CSUR), 52(6),* pp. 1-42. https://doi.org/10.1145/3345317

14. Pudasaini, S., Miralles-Pechuán, L., Lillis, D., & Salvador, M. L. (2024). Survey on Plagiarism Detection in Large Language Models: The Impact of ChatGPT and Gemini on Academic Integrity. https://arxiv.org/pdf/2407.13105

15. Jawahar, G., Abdul-Mageed, M., & Lakshmanan, L. V. (2020). Automatic detection of machine generated text: A critical survey. https://aclanthology.org/2020.coling-main.208

16. Chaplynskyi, D. (2023). Introducing UberText 2.0: A Corpus of Modern Ukrainian at Scale. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pp. 1-10, Dubrovnik, Croatia. Association for Computational Linguistics. https://aclanthology.org/2023.unlp-1.1

17. Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2020). mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. *North American Chapter of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/2021.naacl-main.41

18. GPT-2 Output Detector Demo. (n.d.). https://openai-openai-detector.hf.space/

19. AI Detector – the Original AI Checker for ChatGPT & More. (n.d.). https://gptzero.me/