DOI 10.34132/mspc2025.01.14.01

Катерина Антіпова

АВТОМАТИЗОВАНЕ ВИЯВЛЕННЯ ТЕКСТУ, ЗГЕНЕРОВАНОГО ВЕЛИКИМИ МОВНИМИ МОДЕЛЯМИ

Тези представляють аналіз можливостей сучасних детекторів щодо виявлення тексту згенерованого великими мовними моделями. Зокрема розглянуто використання штучного інтелекту у сфері освіти для академічного шахрайства. Якість згенерованого тексту зростає, що ускладнює виявлення плагіату. Ідентифікація згенерованих робіт та звітів у сфері освіти має деякі відмінності порівняно з іншими сферами в основному через специфіку та контекстуальність навчальних завдань. На теперішній час розроблено багато детекторів для виявлення згенерованого тексту. Сучасний детектор має бути надійним у розпізнаванні таких текстів, щоб запобігти неправомірному використанню штучного інтелекту.

Ключові слова: *академічне шахрайство, великі мовні моделі, генерація тексту, детектори плагіату, штучний інтелект.*

Поява генеративних інструментів штучного інтелекту (ШІ), які базуються на використанні можливостей великих мовних моделей (ВММ), суттєво вплинула на різні галузі людської діяльності. Здатність цих інструментів генерувати текст, схожий на людську мову, становить значну загрозу для академічної доброчесності. Представлення текстів, створених великими мовними моделями, як власної роботи, називається ШІ-плагіатом. Залежність здобувачів вищої освіти від інструментів генерування тексту призводить до втрати креативності та здатності до навчання. Якщо здобувачі отримують оцінки за роботу інших, тоді знижується важлива зовнішня мотивація для отримання знань і компетенцій. Так само спотворюється й оцінка компетентності здобувачів, що може призвести до отримання ними неправомірної вигоди.

Можливості передових ВММ вплинули на сферу освіти різними способами. Наприклад, здобувачі вищої освіти використовують ChatGPT для виконання практичних робіт і складання іспитів. Це викликало занепокоєння щодо поточних систем оцінювання, які використовуються у вищих навчальних закладах. Викладачі намагаються виявляти шахрайські дії здобувачів, і плагіат є однією з головних проблем. У минулому плагіат здебільшого полягав у представленні робіт і есе, які містили абзаци з різних джерел без їх цитування, але з появою ВММ здобувачі тепер можуть використовувати ШІ для генерації текстів для виконання різнопланових завдань.

Типологія плагіату варіюється залежно від типу даних або рівня затемнення. Автори роботи [1] представили різні типології, визначені в кількох дослідницьких роботах, і запропонували нову типологію плагіату за рівнем затемнення: плагіат зі збереженням символів, плагіат зі збереженням синтаксису, плагіат зі збереженням семантики, плагіат зі збереженням ідеї та «гострайтинг» (англ. *ghostwriting*). Тип плагіату може також відрізнятися залежно від типу даних, наприклад, плагіат коду або плагіат тексту. Виявлення плагіату також можна класифікувати на основі кількості мов, використаних в одному тексті. Крім того, коли плагіат виявляється лише за допомогою самого тексту, це внутрішнє виявлення плагіату. Якщо ж плагіат виявляють у порівнянні з іншим текстом, то йдеться про зовнішнє виявлення плагіату.

З передбачуваним швидким розвитком високопродуктивних ВММ якість вихідних текстів зростає, що ускладнює їх виявлення. Навіть людині важко розпізнати згенерований текст через високу здатність ВММ робити його все більше й більше схожим на написаний людиною. Ідентифікація згенерованих робіт та звітів у сфері освіти має деякі відмінності порівняно з іншими сферами в основному через специфіку та контекстуальність навчальних завдань. Хоча загальний контент, створений ШІ, може демонструвати помітні невідповідності в широкому контексті, більш вузький і більш структурований обсяг академічних завдань може маскувати такі аномалії, роблячи процес виявлення більш складним. Отже, виявлення практичних завдань, згенерованих ШІ, вимагає більш досконалих і контекстно-залежних алгоритмів.

Основні відмінності між згенерованим текстом та написаним людиною:

– ШІ дає організовані відповіді з чіткими, структурованими абзацами тексту або пунктами переліку. Текст може виглядати надто відшліфованим або навіть механічним. З іншого боку, текст, написаний людиною, зазвичай більш варіативний по структурі. Може спостерігатися певна дезорганізація або зміна тону, які виникають внаслідок природного ходу думок, через що текст здається менш «досконалим», ніж згенерований.

– ШІ, як правило, використовує нейтральний, майже офіційний тон, якщо немає інших вказівок. Він уникає емоційних крайнощів, на відміну від людини, яка надає емоційні відповіді, наприклад, з доданням певного гумору. Такий текст часто більш динамічний, з різними настроями залежно від ситуації. Загалом, люди, як правило, вносять більше індивідуальності у свої тексти.

– ШІ добре комбінує наявні знання, але він не є по-справжньому творчим у людському розумінні. Його відповіді є похідними, тобто базуються на великій базі знань, і тому приклади можуть не справляти враження свіжих чи оригінальних. Люди набагато кращі у творчому мисленні, часто представляють нові перспективи, несподівані ідеї та унікальні підходи до вирішення проблем.

– ВММ навчені розуміти контекст, але мають певні обмеження при роботі зі складними контекстами певними нюансів. Модель може давати занадто буквальну відповідь, оскільки покладається на узагальнені знання, а не на конкретний особистий досвід.

– У великих згенерованих текстах ШІ, як правило, зберігає зв'язність протягом деякого часу, але може почати втрачати фокус або повторюватися через декілька абзаців. Люди краще зберігають довгострокову зв'язність, але вони також можуть повторювати ідеї або змінювати перспективу.

– Хоча ШІ створює граматично правильний і безпомилковий контент, він все одно може припускатися помилок у міркуваннях або фактах, які людина може легко помітити. Він також може іноді генерувати незграбні фрази, через що текст може здаватися неприродним. Люди частіше роблять граматичні або друкарські помилки, пропускають слова тощо.

Ці узагальнені характеристики вказують на те, що ШІ значно покращився в завданнях обробки природної мови для широкого спектру областей. Порівняно з людиною, йому може бракувати індивідуальності, але він має більш комплексний і нейтральний погляд на питання. Інакше кажучи, згенерований текст має тенденцію бути граматично відшліфованим, структурно чітким і нейтральним за тоном, з тенденцією до точності та узагальнення. З іншого боку, написаний людиною текст має тенденцію до більшої варіативності в тоні, структурі та глибині, з унікальними особистими штрихами та випадковими недосконалостями, які роблять його більш автентичним.

На теперішній час розроблено багато детекторів для виявлення згенерованого тексту, наприклад, DetectGPT [2], RADAR [3], Ghostbuster [4], GPT-Sentinel [5]. Детектор має бути надійним у розпізнаванні текстів, згенерованих ШІ, щоб запобігти неправомірному використанню ВММ. Сучасний детектор контенту, створеного ВММ, повинен мати такі ключові характеристики:

– *Точність* означає, що модель має бути здатною розрізняти згенерований і написаний людиною текст, досягаючи відповідного компромісу між точністю та швидкістю відкликання.

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

– *Узагальнення* означає, що детектор повинен мати можливість працювати узгоджено, незалежно від будь-яких змін в архітектурі моделі, довжини запиту або набору навчальних даних.

– *Пояснюваність* означає, що детектор повинен надавати чіткі пояснення причин своїх рішень. Вона спрямована на те, щоб забезпечити надання зрозумілих для людини міркувань для рішень щодо виявлення, зазвичай представлені у вигляді пояснень природною мовою або візуального представлення основних ознак.

Завдання можна розділити на три рівня:

– пряме пояснення (безпосереднє визначення ознак підробки за допомогою невеликих прикладів у контексті);

– пояснення на основі міркувань (обґрунтування з кількома кроками та оцінка логічної узгодженості);

– детальний аналіз у вільній формі (детальний аналіз аспектів підробки, узгоджений із заздалегідь визначеною таксономією ознак підробки).

Незважаючи на те, що автоматичне виявлення може виявити певний плагіат, існуючі дослідження [6] показали, що програмне забезпечення зіставлення тексту не тільки не знаходить увесь ШІ-плагіат, але й позначає написаний текст як плагіат, таким чином надаючи хибнопозитивні результати. Це неприйнятний сценарій в академічному середовищі, оскільки чесного здобувача можуть звинуватити в порушенні академічної доброчесності.

Сучасні дослідження показують, що найсучасніші детектори ШІ-текстів демонструють значне погіршення продуктивності, коли стикаються з текстами, написаними не носіями англійської мови [7]. Інструменти виявлення також виявляються нездатними впоратися з текстами, перекладеними з інших мов. Крім того, продуктивність детекторів ще більше погіршується з використанням методів обфускації, таких як ручне редагування або машинне перефразування. Згідно зі звітом, опублікованим OpenAI, їхній ШІ-детектор текстів не є повністю надійним у цьому плані. У звіті наводиться оцінка деяких складних випадків для англомовних текстів, їхній класифікатор правильно ідентифікує лише 26% згенерованого тексту, тоді як неправильно класифікує 9% тексту, написаного людиною.

Точність і надійність інструментів виявлення текстів, створених ШІ, може варіюватися залежно від кількох факторів, таких як конкретний інструмент, що використовується, тип моделі ШІ, що генерує текст, і контент, який аналізується. Більшість інструментів виявлення досягають 70-80% точності при виявленні тексту, згенерованого такими моделями, як GPT-3. Детектори також не можуть впоратися з короткими абзацами тексту, а також з результатами роботи більш розвинутих моделей пізніших поколінь, таких як GPT-4.

Отже, проблему академічного шахрайства неможливо вирішити лише за допомогою суто технічних заходів, оскільки така конкуренція може тривати вічно. Таким чином, освітні заклади повинні розглянути можливість прийняття альтернативних освітніх рішень. Одним із рішень для запобігання академічним порушенням може бути вдосконалення поточних стратегій оцінювання в університетах. Оцінювання має більше зосереджуватися на критичному мисленні здобувача та навичках розв'язання проблем, а не на навичках запам'ятовування фактів. Адже прості питання можна легко вирішити за допомогою ВММ. Таким чином, викладачі повинні призначати більш творчі проєкти, створювати екзаменаційні питання з відкритою

відповіддю, які вимагають від здобувачів критичного мислення та застосування їхніх навичок вирішення проблем.

References

1. Foltynek, T., Meuschke, N., & Gipp, B. (2019). Academic plagiarism detection: a systematic literature review. ACM Computing Surveys (CSUR), 52(6), pp. 1-42. URL: https://doi.org/10.1145/3345317
2. 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*. URL: https://doi.org/10.48550/arXiv.2301.11305
3. Hu, X., Chen, P., & Ho, T. (2023). RADAR: Robust AI-Text Detection via Adversarial Learning. URL: https://doi.org/10.48550/arXiv.2307.03838
4. 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*. URL: https://doi.org/10.48550/arXiv.2305.15047
5. Chen, Y., Kang, H., Zhai, V., Li, L., Singh, R., & Ramakrishnan, B. (2023). GPT-Sentinel: Distinguishing Human and ChatGPT Generated Content. URL: https://doi.org/10.48550/arXiv.2305.07969
6. Tang, R., Chuang, Y., & Hu, X. (2023). The Science of Detecting LLM-Generated Text. *Communications of the ACM, 67*, 50-59. URL: https://doi.org/10.48550/arXiv.2303.07205
7. Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J.Y. (2023). GPT detectors are biased against non-native English writers. *Patterns, 4*. DOI: https://doi.org/10.48550/arXiv.2304.02819

DOI 10.34132/mspc2025.01.14.01

Kateryna Antipova

AUTOMATED DETECTION OF TEXT GENERATED BY LARGE LANGUAGE MODELS

The theses present an analysis of the capabilities of modern detectors to detect text generated by large language models. In particular, the use of artificial intelligence for academic fraud is considered. The quality of the generated text is increasing, which makes plagiarism detection more difficult. Identification of generated papers and reports in the field of education has some differences compared to other areas, mainly due to the specificity and contextual nature of educational tasks. Currently there are many instruments developed for detecting generated text. A modern detector should be reliable in recognizing such texts to prevent the misuse of artificial intelligence.

Keywords: *academic fraud, artificial intelligence, large language models, plagiarism detectors, text generation.*

Відомості про автора / Information about the Author

Катерина Антіпова, PhD, доцент (б. в. з.) кафедри інженерії програмного забезпечення факультету комп'ютерних наук Чорноморського національного університету імені Петра Могили, м. Миколаїв, Україна. E-mail: antipova.katerina@chmnu.edu.ua, orcid: https://orcid.org/0000-0002-9012-5290.

Kateryna Antipova, PhD, Associate Professor (w/o a. d.) of the Department of Software Engineering, Faculty of Computer Sciences, Petro Mohyla Black Sea National University, Mykolaiv, Ukraine. E-mail: antipova.katerina@chmnu.edu.ua, orcid: https://orcid.org/0000-0002-9012-5290.

Стаття отримана редакцією 19.05.2025.