

<https://doi.org/10.62405/osiconf.2026.06>

Єфременко А. М.,

Харківська державна академія фізичної культури, Харків, Україна,
ORCID: 0000-0003-0924-0281, e-mail: literasearchukr@gmail.com

Лаєвська Є. Г.,

Харківська державна академія фізичної культури, Харків, Україна,
e-mail: u8764125297@gmail.com

ВИКОРИСТАННЯ ДЕТЕКТОРІВ ШТУЧНОГО ІНТЕЛЕКТУ В ПЕДАГОГІЧНІЙ ДІЯЛЬНОСТІ В СФЕРІ ФІЗИЧНОГО ВИХОВАННЯ ТА СПОРТУ

АНОТАЦІЯ. У статті емпірично досліджено надійність сучасних детекторів штучного інтелекту при перевірці самостійної роботи в текстовому форматі у сфері фізичного виховання та спорту. **Мета.** Емпірично оцінити надійність і узгодженість сучасних детекторів штучного інтелекту при ідентифікації походження виконаних специфічних самостійних завдань у сфері фізичного виховання та спорту. **Методи.** Застосовано метод кейс-стаді. Розроблено три варіанти тренувального комплексу: 1) оригінальний, розроблений науково-педагогічним працівником; 2) згенерований GPTchat; 3) гуманізований засобами Grammarly. Розроблені комплекси вправ перевірено 19 детекторами штучного інтелекту. **Результати.** Встановлено вкрай низьку узгодженість показників (розкид від 0 % до 96 %), відсутність однозначного ефекту гуманізації та ризик хибнопозитивної детекції людського тексту. **Висновки.** Обґрунтовано доцільність оцінювання практико-орієнтованих завдань замість покладання на окремі або декілька автоматизованих детекторів. Перспективним напрямком подальшої розробки даної теми є формування підходу до оцінювання практико-орієнтованих завдань, що поєднує кілька незалежних детекторів та усний захист обґрунтування навчального продукту.

КЛЮЧОВІ СЛОВА: детектори штучного інтелекту; генеративний штучний інтелект; академічна доброчесність; фізичне виховання і спорт.

ВСТУП. Проникнення генеративного штучного інтелекту в освітній процес актуалізує потребу в надійних інструментах верифікації автентичності навчальних текстів. Паралельно з поширенням мовних моделей також зростає ринок комерційних детекторів штучного інтелекту (ШІ) (Perkins, 2023). Їх діагностична точність залишається предметом активної наукової дискусії.

Сучасні дослідження констатують, що результати оцінювання точності детекторів ШІ істотно варіюють залежно від методології тестування, складу вибірки текстів та обраних порогових значень (Hadra, Cambridge & Mesbah, 2026). Частина дослідників зазначає порівняно високу точність окремих комерційних інструментів (зокрема Turnitin), тоді як інші автори фіксують непослідовність показників і системну упередженість детекторів щодо текстів (Liang et al., 2023). Найнадійніші платні детектори демонструють рівень хибнопозитивних результатів на рівні 1–2 %, тоді як безкоштовні та маловідомі сервіси можуть суттєво поступатися їм за точністю (Paulet et al., 2025). У дослідженнях, виконаних на медичних наукових текстах, показано, що усталеного консенсусного порогового значення для класифікації тексту як згенерованого ШІ не існує, а показники в діапазоні 1–20 % визнаються особливо ненадійними (Erol et al., 2025). Окремі публікації прямо застерігають від використання детекторів як єдиної підстави для звинувачення в академічній недоброчесності, наводячи розрахунки, за якими навіть мінімальний рівень похибки в масштабах університету здатен спричинити сотні хибних звинувачень (Shekhar et al., 2026).

Попри значну кількість закордонних досліджень надійності детекторів штучного інтелекту на текстах гуманітарного та академічного профілю, емпіричні дані щодо їх поведінки стосовно вузькос-

спеціалізованих текстів фізкультурно-спортивного спрямування практично відсутні. Це зумовлює потребу в окремій перевірці, оскільки нормативна, таблично структурована форма викладу таких текстів може по-іншому сприйматися алгоритмами детекції порівняно з вільним есеїстичним текстом, на якому побудована більшість наявних досліджень.

Мета — емпірично оцінити надійність і узгодженість сучасних детекторів штучного інтелекту при ідентифікації походження виконаних специфічних самостійних завдань у сфері фізичного виховання та спорту.

МЕТОДИ. Застосовано метод кейс-стаді. Дизайн передбачав створення трьох текстових варіантів одного і того самого навчального завдання. Вони відрізнялися виключно способом їх генерації за незмінності предметного змісту — мети тренування, його структури та переліку вправ.

Емпірична частина дослідження здійснювалася в п'ять послідовних етапів.

На першому етапі тренувальний комплекс було розроблено самостійно, відповідно до самостійного завдання навчальної дисципліни (див. дод. А <https://doi.org/10.5281/zenodo.22305741>). Воно передбачало:

- 1) визначення мети тренування (наприклад, розвиток витривалості, покращення швидкості, загальне зміцнення організму або підготовка до змагань);
- 2) врахування рівня фізичної підготовленості, особи для якої розробляється тренування;
- 3) дотримання традиційної структури заняття фізичними вправами (підготовча, основна та заключна частини);
- 4) з обов'язковим зазначенням типу й змісту вправ, обсягу та інтенсивності навантаження, а також методичних рекомендацій щодо виконання.

Цей варіант тексту позначено як «оригінальний».

На другому етапі ідентичне за змістом завдання було виконано з використанням генеративної мовної моделі GPTchat. Взаємодія з моделлю здійснювалася за принципом рольового промптингу (див. дод. Б <https://doi.org/10.5281/zenodo.22305741>):

- 1) моделі було делеговано роль спортивного тренера, а перший запит відтворював повний текст вихідного завдання та алгоритму його виконання, доповнений вимогою подання комплексу вправ у табличній формі та формування переліку використаних джерел;
- 2) другий, уточнювальний запит слугував для конкретизації цільової аудиторії комплексу та вимоги щодо оформлення джерел відповідно до стандарту ДСТУ з наведенням активних гіперпосилань.

Такий двоетапний промптинг відтворює типову практику ітеративного уточнення завдання, характерну для реального використання генеративного ШІ в навчальній діяльності. Цей варіант тексту позначено як «згенерований».

На третьому етапі згенерований текст було піддано процедурі стилістичної гуманізації («олюднення») із застосуванням сервісу Grammarly. Означене передбачало корекцію синтаксису, лексичного оформлення та стилістики тексту без зміни його змістового наповнення з метою наближення поверхневих текстових характеристик до притаманних людському письму. Цей варіант тексту позначено як «олюднений» (див. дод. В).

На четвертому етапі всі три варіанти тексту (оригінальний, згенерований, олюднений) було перевірено щодо ймовірності їх створення ШІ за допомогою 19 незалежних програмних детекторів.

Застосовані сервіси суттєво відрізняються за форматом вихідного показника. Частина з них (Grammarly, Plag, Justdone, Humbot/humalingo, Getmerlin, Plagiarismsearch, Plagiarismdetector, Switas, Openl, Sidekicker, Gptzero, Sapling, Quillbot.) повідомляють безпосередньо ймовірність ШІ-генерації

тексту, тоді як інші (Isgen, Plagium, Duplichecker, Plagly, Gowinston, Paraphrasetool) оперують оберненим показником — ймовірністю людського авторства. Для забезпечення порівнюваності результатів здійснено гармонізацію показників. Так, значення другої групи сервісів перераховано за формулою:

Гармонізований показник:

$$100 \% \text{ — вихідний показник людського авторства} \quad (1)$$

Внаслідок цього отримано уніфікований показник — гармонізовану ймовірність ШІ-генерації.

Статистичну обробку отриманих даних виконано методами описової статистики з розрахунком для кожного варіанта тексту: середнього значення; стандартного відхилення; розмаху варіювання.

РЕЗУЛЬТАТИ ТА ОБГОВОРЕННЯ. Реалізація трьох варіантів тренувального комплексу дозволила отримати текстовий матеріал, ідентичний за предметним змістом (мета тренування, структура заняття, перелік вправ) і диференційований виключно за способом створення. Генерація тексту засобами GPTchat за рольовим промптом «спортивний тренер» з уточненням щодо цільової аудиторії дозволила отримати структурно повноцінний комплекс у табличній формі, що відповідав усім вимогам вихідного завдання. Наступна гуманізація засобами Grammarly змінила стилістичні та синтаксичні характеристики тексту без втручання в його змістове наповнення.

Перевірка трьох отриманих варіантів детекторами ШІ виявила показники, що суттєво відрізнялися за форматом (таблиця 1):

Таблиця 1. Результати перевірки текстів детекторами ШІ (вихідні показники сервісів)

Детектор	Згенерований (GPTchat)	Олюднений (Grammarly)	Оригінальний
Grammarly	0 % (ШІ)	—	0 % (ШІ)
Plag	30 % (ШІ, низький ризик)	41 % (ШІ, середній ризик)	—
Justdone	88 % (ШІ)	88 % (ШІ)	88 % (ШІ)
Humbot / humalingo	83 % (ШІ)	94 % (ШІ)	80 % (ШІ)
Isgen	100 % (людина)	100 % (людина)	—
Getmerlin	48 % (ШІ)	45 % (ШІ)	0 % (ШІ)
Plagiarismsearch	48 % (ШІ)	44 % (ШІ)	—
Plagium	100 % (людина)	100 % (людина)	—
Duplichecker	100 % (людина)	100 % (людина)	—
Plagiarismdetector	78 % (ШІ)	78 % (ШІ)	—
Plagly	68 % (людина)	68 % (людина)	—
Switas	4 % (ШІ)	43 % (ШІ)	—
Paraphrasetool	100 % (людина)	100 % (людина)	—
Openl	48 % (ШІ)	45 % (ШІ)	—
Sidekicker	0–20 % (ШІ)	0–20 % (ШІ)	—
Gowinston	96 % (людина)	95 % (людина)	—
Gptzero	64 % (ШІ)	30 % (ШІ)	0 % (ШІ)
Sapling	42,2 % (ШІ)	48,9 % (ШІ)	25,6 % (ШІ)
Quillbot	96 % (ШІ)	41 % (ШІ)	0 % (ШІ)

13 сервісів повідомляли безпосередньо ймовірність ШІ-генерації;

6 сервісів оперували оберненим показником — ймовірністю людського авторства.

Для забезпечення порівнюваності результатів усі показники другої групи гармонізовано за формулою «100 % мінус вихідний показник людського авторства» до єдиної шкали ймовірності ІІІ-генерації (таблиця 2).

Таблиця 2. Гармонізована ймовірність ІІІ-генерації тексту за версіями (%)

Детектор	Згенерований, % ІІІ	Олюднений, % ІІІ	Оригінальний, % ІІІ
Grammarly	–	–	–
Plag	30,00	41,00	–
Justdone	88,00	88,00	88,00
Humbot / humalingo	83,00	94,00	80,00
Isgen	–	–	–
Getmerlin	48,00	45,00	–
Plagiarismsearch	48,00	44,00	–
Plagium	–	–	–
Duplichecker	–	–	–
Plagiarismdetector	78,00	78,00	–
Plagly	32,00	32,00	–
Switas	4,00	43,00	–
Paraphrasetool	–	–	–
Openl	48,00	45,00	–
Sidekicker	10,00	10,00	–
Gowinston	4,00	5,00	–
Gptzero	64,00	30,00	–
Sapling	42,20	48,90	25,60
Quillbot	96,00	41,00	–

Аналіз наведених у таблиці 1 даних засвідчує вкрай високу міжсервісну варіативність оцінок. Для згенерованого варіанта тексту розкид гармонізованих значень становив від 0 % (Grammarly, Isgen, Plagium, Duplichecker, Paraphrasetool) до 96 % (Quillbot), тобто практично повний діапазон шкали.

Середнє значення ймовірності ІІІ-генерації для згенерованого варіанта ($M = 35,50$ %; $SD = 33,10$ %) практично не відрізняється від середнього значення для гуманізованого варіанта ($M = 35,80$ %; $SD = 29,10$ %). Означене суперечить очікуваному ефекту олюднення тексту як засобу зниження його детектованості. Коефіцієнт варіації в усіх трьох групах перевищував 90 %. Означене вказує на надзвичайно низьку конвергентну валідність між застосованими детекторами. Такий висновок узгоджується з наведеними вище даними закордонних досліджень щодо непослідовності комерційних інструментів детекції (Chaka, 2024; Elkhatat, Elsaid & Almeer, 2023).

Розгляд динаміки показників окремих сервісів після гуманізації тексту виявив різноспрямовані тенденції. У частини детекторів гуманізація призвела до зростання показника ймовірності ІІІ-генерації: Plag (з 30,00 % до 41,00 %); Humbot/humalingo (з 83,00 % до 94,00 %); Switas (з 4,00 % до 43,00 %); Sapling (з 42,20 % до 48,90 %). В іншій частині сервісів, навпаки, зафіксовано зниження показника: Gptzero (з 64,00 % до 30,00 %); Quillbot (з 96,00 % до 41,00 %), Getmerlin, Openl та Plagiarismsearch (з 48,00 % до 44–45,00 %). У решти сервісів показник практично не змінився. Такий різноспрямований характер змін підтверджує тезу дослідників про те, що ефективність паразитичних технік «обходу» детекторів (зокрема парафразування та стилістичного редагування) є вибірковою та залежною від конкретної архітектури кожного окремого сервісу (Walters, 2023).

Окремої уваги заслуговують результати перевірки оригінального (повністю людського) варіанта тексту. Попри те що цей текст був створений виключно людиною, 3 з 7 детекторів, що його перевіряли, зафіксували істотну ненульову ймовірність ШІ-генерації: Justdone — 88,00 %; Humbot/humalingo — 80,00 %; Sapling — 25,60 %. Означене узгоджується з наведеними в огляді літератури даними про підвищений ризик хибнопозитивних результатів для нормативно-структурованих текстів і текстів осіб із нетиповими для навчальних вибірок детекторів мовностилістичними особливостями (Perkins Furze & Roe, 2024; Weber-Wulff et al., 2023).

Отримані результати дозволяють виявити суттєву розбіжність між технологічними очікуваннями та педагогічною реальністю. Стрімке впровадження генеративного штучного інтелекту зумовило інституційну відповідь у формі масового застосування детекторів, проте ефективність такого підходу в підготовці майбутніх тренерів та вчителів фізичного виховання залишається дискусійним питанням.

З огляду на викладене, доцільним є формування комплексного підходу до оцінювання академічної доброчесності виконання практико-орієнтованих завдань у сфері фізичного виховання та спорту. Моделлю може виступати використання кількох незалежних джерел доказовості:

- 1) використання не менше двох-трьох детекторів ШІ з різною архітектурою;
- 2) аналіз відеофіксації безпосереднього виконання тренувального комплексу як незамінного доказу практичної (рухової) компетентності здобувача;
- 3) усне обговорення (захист) обґрунтування побудови комплексу, яке складно відтворити виключно засобами генеративного ШІ без глибокого розуміння предметної галузі.

Перспективи подальших досліджень пов'язані з розширенням емпіричної бази за рахунок більшої кількості навчальних завдань і незалежних виконавців, а також розробкою галузевих рекомендацій щодо етичного і педагогічно виваженого використання генеративного штучного інтелекту та засобів його детекції в підготовці фахівців фізичної культури і спорту.

Обмеження. Методика дослідження має ряд обмежень, урахованих під час інтерпретації результатів. Так, оригінальний варіант тексту було перевірено лише 7 детекторами, що зумовлює меншу статистичну потужність порівнянь за його участю. Даний підхід було застосовано свідомо, вибірково здійснюючи перевірку сервісами, які викликали суперечливі або, навпаки, релевантні результати оцінювання тренувань генерованого та олюдненого. Вибірка текстів обмежена одним випадком ($n = 1$ навчальне завдання в трьох варіантах виконання), що не дозволяє узагальнювати отримані кількісні значення на генеральну сукупність навчальних текстів. Але, це й не применшує цінності виявлених якісних закономірностей. Алгоритми та порогові значення переважної більшості комерційних детекторів є проприєтарними та не розкриваються розробниками, а самі моделі підлягають регулярному оновленню, що обмежує відтворюваність результатів у часі. Для сервісу Justdone в усіх трьох варіантах тексту зафіксовано ідентичний показник (88 %) з однаковою структурою деталізації, що є нетиповим для текстів різного походження і потребує окремого критичного розгляду як методологічна аномалія (можливий технічний збій, кешування результату або нечутливість алгоритму до входних відмінностей).

ВИСНОВКИ. Проведене емпіричне дослідження засвідчило, що застосовані детектори штучного інтелекту демонструють вкрай низьку взаємну узгодженість при перевірці виконаних самостійних завдань фізкультурно-спортивного спрямування. Встановлено, що стилістична гуманізація тексту не має однозначного, стабільного ефекту зниження його детектованості. Так, у рівній частині перевірених сервісів показник ймовірності ШІ-генерації після олюднення, навпаки, зростає. Виявлено суттєвий ризик хибнопозитивної детекції повністю людського тексту, що має нормативну, табличноструктуровану форму викладу, притаманну методичним матеріалам зі спортивного тренування. Отримані результати узгоджуються із загальною тенденцією і доводять непридатність окремого автоматизованого детектора як самодостатнього критерію встановлення академічної недоброчесності.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

- Chaka, C. (2024). Reviewing the performance of AI detection tools in differentiating between AI-generated and human-written texts. *Journal of Applied Learning & Teaching*, 7(1), 115–126. <https://doi.org/10.37074/jalt.2024.7.1.25>
- Elkhatat, A. M., Elsaid, K., & Almeer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity*, 19(1), Article 17. <https://doi.org/10.1007/s40979-023-00140-5>
- Erol, G., Ergen, A., Gülşen Erol, B., Kaya Ergen, Ş., Bora, T. S., Çölgeçen, A. D., & Güngör, A. (2025). Can we trust academic AI detective? Accuracy and limitations of AI-output detectors. *Acta Neurochirurgica*, 167(1), Article 214. <https://doi.org/10.1007/s00701-025-06622-4>
- Hadra, M., Cambridge, K., & Mesbah, M. (2026). Evaluating the accuracy and reliability of AI content detectors in academic contexts. *International Journal for Educational Integrity*, 22(4), 1–18. <https://doi.org/10.1007/s40979-026-00213-1>
- Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, 4(7), Article 100779. <https://doi.org/10.1016/j.patter.2023.100779>
- Paullet, K., Pinchot, J., Kinney, E., & Stewart, T. (2025). Precision check: A critical look at the reliability of AI detection tools. *Issues in Information Systems*, 26(3), 400–412. https://doi.org/10.48009/3_iis_2025_2025_132
- Perkins, M. (2023). Academic integrity considerations of AI large language models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching & Learning Practice*, 20(2), 1–24. <https://doi.org/10.53761/1.20.02.07>
- Perkins, M., Furze, L., & Roe, J. (2024). Simple techniques to bypass GenAI text detectors: Implications for inclusive education. *International Journal of Educational Technology in Higher Education*, 21(1), Article 53. <https://doi.org/10.1186/s41239-024-00487-w>
- Shekhar, A. R., Agarwal, S., Bordoloi, P., Shah, Y., Anvekar, T., & Gupta, V. (2026). DoPE: Decoy Oriented Perturbation Encapsulation Human-Readable, AI-Hostile Documents for Academic Integrity (arXiv:2601.12505). *arXiv*. <https://doi.org/10.48550/arXiv.2601.12505>
- Walters, W. H. (2023). The effectiveness of software designed to detect AI-generated writing: A comparison of 16 AI text detectors. *Open Information Science*, 7(1), Article 20220158. <https://doi.org/10.1515/opis-2022-0158>
- Weber-Wulff, D., Anohina-Naumeca, A., & Bjelobaba, S. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19(1), 1–39. <https://doi.org/10.1007/s40979-023-00146-z>

Yefremenko A. M.,

Kharkiv state academy of physical culture, Kharkiv, Ukraine,
ORCID: 0000-0003-0924-0281, e-mail: literasearchukr@gmail.com

Laievska Ye. H.,

Kharkiv state academy of physical culture, Kharkiv, Ukraine,
e-mail: u8764125297@gmail.com

THE USE OF AI DETECTORS IN PEDAGOGICAL PRACTICE WITHIN PHYSICAL EDUCATION AND SPORTS

ABSTRACT. *To empirically assess the reliability and consistency of modern artificial intelligence detectors when identifying the origin of completed specific independent tasks in the field of physical education and sports. **Methods.** The case study method was applied. Three versions of the training complex were developed — the original*

one, generated by GPTchat, and humanized using Grammarly tools, which were tested by 19 AI detectors. **Results.** Extremely low consistency of indicators was established (variation from 0% to 96%), the absence of an unambiguous humanization effect and the risk of false positive detection of human text. **Conclusions.** The feasibility of assessing practice-oriented tasks instead of relying on a separate or several automated detectors was substantiated. A promising direction for further development of this topic is the formation of an approach to the assessment of practice-oriented tasks, combining several independent detectors and oral defense of the justification of the educational product.

KEYWORDS: artificial intelligence detectors; generative artificial intelligence; academic integrity; physical education and sports.