

**Оценка на способността за разпознаване на AI генериран текст
чрез емпирично изследване**
Дилян Димитранов

**Assessment of the Ability to Recognize AI-Generated Text
Through Empirical Research**
Dilyan Dimitranov

Abstract:

This report investigates the ability of students and faculty to distinguish between text generated by AI language models (Chat GPT 3.5 and Gemini 1.5) and text written by academic staff. An empirical study was conducted among 110 students and 20 faculty members who analyzed texts on six different scientific topics (Economics, Astronomy, Meteorology, Scientific Publications, Emergency Situations, Environment) in Bulgarian and English. The results demonstrate significant difficulties for both groups in identifying the origin of the text. The main conclusion is that the boundary between AI-generated and human text is increasingly blurring, which calls into question traditional assessment methods and requires a rethinking of approaches in the digital transformation of education.

The findings highlight the need to develop adequate strategies and tools to address the challenges posed by the advancement of AI technologies in the educational sphere.

Keywords: AI generated text; text identification; empirical study; digital transformation; LLM; academic integrity.

For contacts: Dilyan Dimitranov, PhD, NVNA “N. Vaptsarov”, d.dimitranov@nvna.eu

ВЪВЕДЕНИЕ

Бързото развитие и широкото навлизане на изкуствения интелект (AI), особено на големите езикови модели (LLMs) като ChatGPT и Gemini, променят редица аспекти на съвременния живот, включително и образованието. Тези инструменти предлагат нови възможности за достъп до информация, автоматизация на рутинни задачи и дори подкрепа в организацията на учебния процес. Паралелно с това обаче, способността им да генерират кохерентен, стилистично издържан и често правдоподобен текст, който може да имитира човешко писане, поставя сериозни предизвикателства пред академичната общност, автентичността на студентските разработки както и ефективността на традиционните методи за оценка на проекти във висшето образование.

В контекста на ускорената дигитална трансформация, на която е подложено образованието през последните години, възниква критичният въпрос доколко студенти и преподаватели – са в състояние адекватно да разпознават текст, създаден от AI, в сравнение с текст, написан от човек (напр. от академичен състав). Тази способност е от ключово значение не само за предотвратяване на академични нарушения като плагиатство, но и за адаптирането на педагогическите подходи и инструментите за оценка към новата реалност, в която AI е леснодостъпен и широко използван.

Настоящото изследване поставя за цел, да оцени способността на студенти и преподаватели във висше училище да различават текст, генериран от популярни AI модели, от академичен текст, взет от утвърдени учебни материали. Чрез провеждане на контролирана анкета, изследването се стреми да даде представа

за актуалното ниво на "AI текстова грамотност" и да предостави емпирични данни, които да послужат за основа относно необходимите промени и адаптации в образователната система в условията на напредваща дигитализация и интеграция на AI.

ЛИТЕРАТУРЕН ОБЗОР

Бързото развитие на изкуствения интелект, особено при големите езикови модели като ChatGPT, води до създаването на инструменти, способни да генерират текст с високо качество [1]; [4]. Тази възможност поставя сериозни предизвикателства пред академичната почтеност и традиционните методи за оценка във висшето образование [3]; [5]. Установено е, че текст, генериран от LLMs, може да бъде труден за разпознаване както от съществуващи технологични средства за детекция, така и от обучен академичен персонал [4].

Въпреки че AI инструментите като ChatGPT могат да бъдат предразположени към грешки и фабрикуване на информация [1], способността им да създават кохерентен текст налага преосмисляне на нормите [1] и адаптиране на педагогическите стратегии и оценяването [2]; [5]. Академичната общност ще трябва да премине през етапи на адаптиране на политиките за академична почтеност [3]; [4]. В този контекст, изследването на човешката способност за разграничаване на AI генериран текст от ръчно писан текст е ключово за разбирането на предизвикателствата и ефективното адаптиране на образователния процес към реалността на широкодостъпния AI.

МЕТОДОЛОГИЯ

Настоящото изследване се основава на емпиричен подход, осъществен чрез провеждане на онлайн анкета. В проучването взимат участие общо 130 души, разпределени в две групи: 110 студента и 20 преподавателя от ВВМУ “Н. Й. Вапцаров”. Участието в анкетата е доброволно. В началото на анкетния формуляр са събрани демографски данни (възрастови диапазони, статус студент/преподавател) и информация относно опита на участниците с големи езикови модели, включително конкретни използвани модели.

Анкетата включва поредица от задачи за разпознаване на произхода на текст. Участниците анализират текстове по осем различни тематични категории на български и английски език, сред които Астрономия, Околна среда, Аварийни ситуации, Икономика, Научни публикации и Метеорология. Задачите са структурирани по различен начин: сравнение на двойки текстове (един генериран от AI, един човешки), анализ на групи от три или четири текста с различен произход, или идентифициране на произхода на отделни абзаци в рамките на един текст. Текстовете, написани от човек, представляват директно копирани откъси от академични публикации (учебници, статии, стандарти). Текстовете, генерирани от AI, са създадени посредством моделите Chat GPT 3.5 и Gemini (свободна версия към август 2024 г.), като за всяка тема са използвани идентични подкани към двата модела. Редът на представяне на текстовете във всяка задача е разбъркан за всеки участник с цел намаляване на ефекта на последователност.

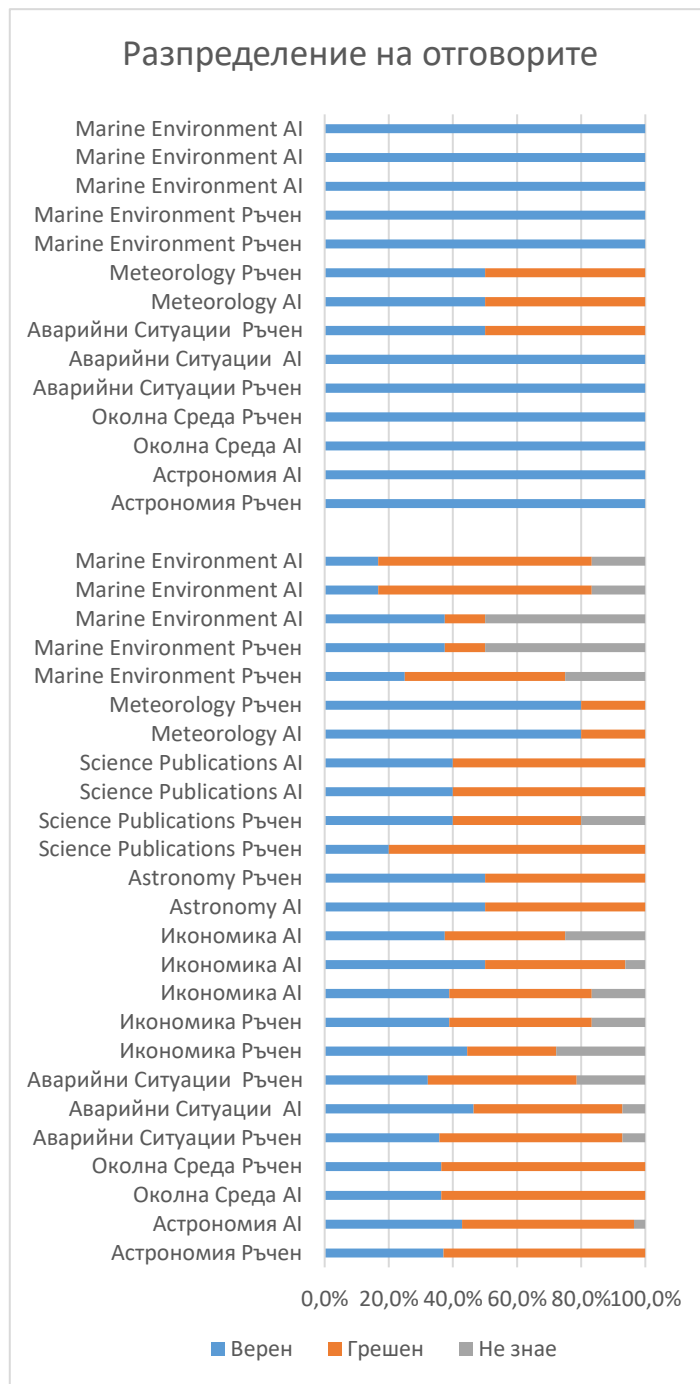
Събраните данни от анкетата са експортирани от платформата Google Forms и подготвени за статистически анализ. Предварителният анализ включва разделяне на резултатите по групи участници (студенти и преподаватели) за да се оцени способността им за правилно идентифициране на произхода на текста.

РЕЗУЛТАТИ

Анализът на данните от проведеното емпирично изследване разкрива значителни предизвикателства пред участниците при разпознаването на произхода на текст в академичен контекст. Резултатите се анализират въз основа на разпределението на отговорите ("AI текст", "Ръчно написан текст" и "Не мога да определя") за всеки представен текст, категоризирани по тематични области и групи анкетирани.

В обобщение, както студентите, така и преподавателите демонстрират затруднения при надеждното разграничаване на текст, генериран от големи езикови модели, от текст, създаден от човек. Високият процент на погрешни идентификации е характерен за повечето тематични категории и при двете изследвани групи. Наличието на отговори "не мога да определя" в значителен дял от случаите допълнително подчертава несигурността на анкетираните при определяне на произхода на текста.

Резултати по групи участници
 При студентите се наблюдава устойчива тенденция към неточно идентифициране на произхода на текстовете в повечето изследвани категории. Например, при ръчно



са отговори на преподаватели, а във втората – на студенти.

Фиг. 1. Общо разпределение на отговорите по тип текст и група участници. В първата секция

написания текст по Астрономия (на Български), 63% от студентите погрешно го класифицират като AI генериран, докато при AI текста по същата тема 53.6% погрешно го определят като ръчно написан.

Сходни ниски нива на правилно разпознаване се забелязват в категории като Околна среда (на Български) и Научни публикации (на Английски), където дялът на погрешните отговори често надхвърля 60%. Дори в категории с относително по-висока успеваемост, като Метеорология (на Английски), където 80% от студентите правилно разпознават AI текста, все още е налице значителен процент грешки при ръчно написания текст (20% погрешно сметнат за AI).

При преподавателите картината е по-нееднозначна. В някои категории като Астрономия (на Български), Околна среда (на Български) и Морска среда (на Английски), преподавателите демонстрират 100% правилно разпознаване на произхода на текстовете. Това предполага, че в области с висока степен на специализация или в текстове с ясно изразени стилови характеристики, опитът играе ключова роля. Въпреки това, в други категории, резултатите при преподавателите са сходни със затрудненията при студентите (50% погрешни идентификации в Метеорология). В секцията Аварийни ситуации (Български), вторият ръчно написан текст е погрешно определен като AI от 50% от преподавателите.

ЗАКЛЮЧЕНИЕ

Настоящото емпирично изследване оценява способността на студенти и преподаватели във висше училище да разграничават текст, генериран от AI езикови модели (Chat GPT 3.5 и Gemini), от текста, написан от академичен състав. Резултатите показват, че участниците изпитват значителни затруднения при надеждното идентифициране на произхода на текста. Високият процент на погрешни класификации и наличието на отговори "не мога да определя" предполагат липсата на универсални критерии за визуално разграничаване между двата типа текст. Опитът в академичното писане се оказва е недостатъчен за надеждно откриване на AI генериран текст, особено когато моделите имитират научен стил.

Предвид установените затруднения при разпознаването на AI генериран текст и широкото навлизане на AI инструменти сред студентите, настоящото изследване препоръчва преосмисляне на педагогическите практики и методи за оценяване. Специфични мерки включват намаляване на обема на писмените домашни работи, засилване контрола по време на изпити и фокусиране върху задачи с практическа насоченост или тестови формати. За преподавателите се налага развиване на повишена критичност към студентските проекти.

Въпреки ограниченията на настоящото проучване, проведено в специфичен морски контекст с извадка от морско училище, резултатите подчертават че преподавателите трябва да са наясно с трудностите при разграничаване на студентски текстове от AI генерирани такива. Бъдещи изследвания могат да оценят ефективността на специализирани LLM детектори в образователна среда.

Този доклад е финансиран от Министерството на образованието и науката по Национална научна програма "Сигурност и отбрана", осъществявана в изпълнение на Националната стратегия за развитие на научните изследвания 2017-2030 и приета с Решение на Министерски съвет № 731 от 21 октомври 2021 г. Материалът отразява единствено мнението на автора и МОН не носи отговорност за съдържанието.

ЛИТЕРАТУРА

- [1] Currie, G. M. (2023). Academic integrity and artificial intelligence: Is ChatGPT hype, hero or heresy? *Seminars in Nuclear Medicine*, 53(5), 719–730. <https://doi.org/10.1053/j.semnuclmed.2023.04.008>
- [2] Komitov, D. (2024). Issues and concerns in the maritime education regarding cyber risk management. *International scientific journal "trans & motauto world"*, 9(2), 63–65.
- [3] Kumar, R., Eaton, S. E., Mindzak, M., & Morrison, R. (2024). Academic integrity and artificial intelligence: An overview. In S. E. Eaton (Ed.), *Second handbook of academic integrity*. Springer. https://doi.org/10.1007/978-3-031-54144-5_153
- [4] Perkins, M. (2023). Academic integrity considerations of AI large language models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching and Learning Practice*, 20(2), 1–24. <https://search.informit.org/doi/10.3316/informit.T2024111300009591751711095>
- [5] Sullivan, M., Kelly, A., & McLaughlan, P. (2023). ChatGPT in higher education: Considerations for academic integrity and student learning. *Journal of Applied Learning & Teaching*, 6(1), 1–10. <https://doi.org/10.37074/jalt.2023.6.1.17>