Между интуицията и анализа: студентски стратегии за идентифициране на AI генерирани текстове

Владислав Маринов, Анита Тодоранова

Between Intuition and Analysis: Student Strategies for Identifying AI-Generated Texts

Vladislav Marinov, Anita Todoranova

Abstract:
This paper presents an experimental study examining the ability of two groups of students to identify whether a given text was generated by artificial intelligence (AI) or authored by a human, in the context of the increasing use of generative models in education and digital communication. During the experiment, participants indicated the presumed origin of the text, as well as their level of confidence and the arguments supporting their decision.

Keywords: artificial intelligence, generative models, machine-generated text, textual authenticity, critical text analysis.

For contacts: For contacts: Asst. Prof. A. Todoranova, St. Cyril and Methodius University of Veliko Tarnovo, a.todoranova@ts.uni-vt.bg

The great opportunity offered by ICTs comes with a huge intellectual responsibility to understand them and take advantage of them in the right way.
Luciano Floridi

ВЪВЕДЕНИЕ

Развитието на генеративните модели на изкуствен интелект (AI) доведе до съществена промяна в начина, по който се създават, разпространяват и възприемат текстове в дигитална среда. Съвременните големи езикови модели (LLM), като ChatGPT, Claude и Gemini и др., са способни да генерират текстове с висока степен на граматична коректност, структурираност и езикова последователност, които в редица случаи са трудно различими от „човешкото“ писане. Тази технологична еволюция не само разширява възможностите за автоматизирано създаване на съдържание, но и поставя под въпрос установените представи за авторство, оригиналност и езикова автентичност.

В контекста на образованието навлизането на LLM поражда редица предизвикателства, свързани с преценката за степента на оригиналност на студентски изпитни работи. Традиционните критерии за „качествен текст“ – яснота, логическа свързаност, граматична правилност и аргументативна последователност – все по-често се оказват недостатъчни за разграничаване между „човешко“ и „машинно“ създадено съдържание.

Също така релевантен е въпросът доколко реципиентите – и в частност студентите – могат надеждно да разпознават AI генерирани текстове и въз основа на какви критерии се извършва преценката. Изследванията в тази област показват, че както специализираният софтуер за откриване на AI генерирани текстове (AI content detection software), така и човешките оценители често разчитат на интуитивни маркери и предварителни очаквания (или алгоритми при софтуера), които невинаги кореспондират с реалните езикови характеристики на текста.

Настоящият доклад се фокусира върху идентифицирането от студенти на AI генериран текст и върху аргументите им за определяне на произхода му. Изследването се стреми не само да установи степента на разпознаваемост на текст, създаден от LLM, но и да анализира когнитивните и езиковите фактори, които влияят върху процеса на оценяване. В този смисъл настоящата работа се вписва в по-широкия дебат за ролята на изкуствения интелект в образователната среда – не само като предизвикателство, но и като потенциален инструмент за подпомагане на обучението и предоставяне на автоматизирана, насочваща обратна връзка.

ИЗЛОЖЕНИЕ

Целта на настоящия доклад е да се изследва доколко студентите могат надеждно да разпознават AI генериран текст, както и да се проследи кои езикови, стилови и съдържателни характеристики влияят върху техните оценки. Анализът се фокусира не само върху крайния избор („AI“ или „човек“), но и върху аргументационните стратегии, които мотивират избора.

В изследването участват 30 студенти от Великотърновския университет „Св. св. Кирил и Методий“ от специалностите „Българска филология“ (IV курс) и „Софтуерно инженерство“ (I курс). Подборът на участниците в изследваните групи позволява сравнение между студенти с хуманитарен или с технически профил, при които текстът има различна функционална значимост в обучението и професионалната им реализация.

Дизайнът на експеримента включва кратък текст, генериран от AI, и въпроси, свързани с оценка на яснотата, логическата свързаност, граматичната коректност, оригиналността, емоционалния тон и предполагаемия произход на текста. Също така респондентите отбелязват и степента си на увереност при направения избор.

Резултатите от анкетното проучване показват, че по отношение на яснотата, логическата свързаност и граматичната коректност анализираният текст получава преобладаващо високи оценки и от двете групи респонденти. Това потвърждава, че AI генерираният текст отговаря на утвърдените представи за „качествен академичен текст“ по отношение на повърхностния (формален) аспект. Подобни резултати кореспондират с наблюденията на Gehrmann et al. (2019), според които съвременните генеративни модели успешно възпроизвеждат нормативните характеристики на академичното писане. Добрата структура и малкото на брой езикови и стилистични грешки често се използват от участниците в експеримента като аргумент в полза на „изкуствения“ произход на текста, което разкрива наличието на устойчив когнитивен шаблон, при който „прекалено правилното“ и „прекалено подреденото“ се възприемат като неавтентични. Подобен парадокс е описан и от Clark et al. (2021), които отбелязват, че текстове, следващи твърде стриктно нормативните модели, често пораждат съмнение относно „човешкото“ им авторство.

Оценките за оригиналност и емоционален тон се оказват значително по-колебливи и варират в по-широк диапазон. Част от студентите интерпретират неутралния и балансиран стил като липса на личностна позиция или субективна ангажираност, което автоматично асоциират с AI. Други участници, напротив, възприемат същите характеристики като типични за академичното писане, при

което личната експресивност е умишлено редуцирана. Това разминаване показва, че респондентите често смесват жанровите очаквания (как трябва да изглежда академичният текст) с очакванията за авторство (как би го написал човекът или машината).

В този смисъл анкетата разкрива не толкова обективна способност за разпознаване на AI генериран текст, колкото конфликт между различни представи за стил, норма и автентичност. Резултатите потвърждават тезата на Sadasivan et al. (2023), че както човешките оценители, така и специализираният софтуер за откриване на AI генерирани текстове срещат сериозни затруднения при разграничаването на текстове, които са граматически коректни, логически последователни и стилистично неутрални.

Съществен и особено показателен резултат е несъответствието между декларираната увереност в отговорите и качеството на аргументацията. Значителна част от респондентите изразяват висока степен на увереност при определяне на произхода на текста, но не успяват да формулират конкретни аргументи при отговора на въпроса: „Какво Ви насочи към този извод?“, което показва, че причина за увереността по-често е интуитивното усещане, а не аналитичният процес. Чрез проведената анкета индиректно се измерва не само идентификацията на AI генериран текст, но и нивото на метакогнитивна осъзнатост при работа с писмено съдържание.

Анализът по специалности разкрива ясно диференцирани аргументационни стратегии. Студентите от специалност „Българска филология“ демонстрират по-развита метаезикова рефлексия, като по-често посочват стилови повторения, синтактична еднотипност или липса на експресивност. В същото време при тях се наблюдава по-висока степен на очакване, че „човешкият“ текст трябва да съдържа индивидуални отклонения от нормата. Това потвърждава вече отбелязаното наблюдение, че високата езикова коректност се интерпретира като сигнал за текст, генериран от AI. От друга страна, студентите от специалност „Софтуерно инженерство“ прилагат по-прагматичен и функционален модел на оценка. Техните аргументи се фокусират върху това дали текстът е ясен, логичен и „работещ“ като средство за предаване на информация. Анализът на езиковите средства остава вторичен, което често води до по-интуитивен избор. Тези различия кореспондират с наблюденията на Weber-Wulff et al. (2023) относно това, че даден текст се разбира, оценява и използва в рамките на конкретна област (discipline-specific text perception) и специфичните критерии за оценка в различни академични общности.

Последният елемент на анкетата, в който респондентите трябваше да мотивират избора си по отношение на авторството на текста, се оказа най-информативният, тъй като позволи ясно разграничаване между интуитивните отговори, описателните, но повърхностни аргументи, и аналитичните, рефлексивни коментари. Това е и основният принос на изследването – беше установено, че с по-висока степен на релевантност е начинът, по който се мотивира предполагаемото авторство, а не бинарният избор „AI или човек“.

Фигура 1. Обобщен модел на експеримента

ЗАКЛЮЧЕНИЕ

Настоящото експериментално изследване показва, че разпознаването на AI генерирани текстове от страна на студентите е ограничено и силно зависимо от предварителни очаквания, интуитивни стратегии и жанрови стереотипи, независимо от образователния профил. Високата степен на граматична коректност и логическа подреденост на текста често се възприема като индикатор за „изкуствен“ произход, което разкрива парадокс в съвременното възприемане на академичното писане – нормативността, традиционно считана за белег на качество, се интерпретира като признак за текст, генериран от AI.

Сравнителният анализ между специалностите обаче показва, че макар респондентите често да достигат до сходни крайни оценки относно произхода на текста, те използват различни аргументационни стратегии. Студентите от специалност „Българска филология“ демонстрират по-развита метаезикова рефлексия и по-често аргументират избора си чрез стилови, дискурсивни и синтактични характеристики. При тях се наблюдава ясно изразено очакване, че „човешкият“ текст следва да съдържа индивидуални отклонения, експресивност или вариативност, които отсъстват в създадения от AI текст. Докато студентите от специалност „Софтуерно инженерство“ прилагат по-прагматичен и функционален модел на оценка, фокусиран върху яснота, логичност и ефективност на текста като средство за комуникация. Техният избор по-често се основава на цялостно

интуитивно възприемане, при което анализът на езиковата форма и стилистичните детайли остава вторичен. Това води до различен тип аргументация, която е по-малко аналитична, но не и по-малко последователна в рамките на използваните критерии.

Тези различия показват, че изследването не разкрива толкова разлика в способността за разпознаване на AI генерирани текстове, колкото в когнитивните стратегии, чрез които студентите осмислят текста и вземат решение. В по-широк план настоящото изследване е не само анализ на нагласите за ролята на изкуствения интелект, но и инструмент за по-добро разбиране на когнитивните и образователните ефекти от навлизането на генеративния AI в академичната среда. Резултатите показват, че взаимодействието с AI генерирани текстове може да бъде използвано като диагностично средство за оценка на критичното мислене, аргументативните умения и метатекстовата осъзнатост на студентите от различни специалности.

От педагогическа гледна точка резултатите подкрепят необходимостта от преосмисляне на образователните практики в условията на широко разпространени генеративни AI технологии, тъй като ефективното функциониране на сложните системи изисква адаптация, иновации и устойчивост (Byanov, 2025). Както отбелязва Л. Тодоранова, „интегрирането на изкуствен интелект в обучението съответства на съвременните подходи като компетентностно базираното обучение (CBT), което акцентира върху изграждането на практически умения чрез решаване на реални проблеми“ (Тодоранова, 2025: 536). Вместо да се фокусира върху „разобличаването“ на AI генерирани текстове, обучението следва да насърчава развитие на умения за критичен текстов анализ, аргументация и метарефлексия. В този контекст изкуственият интелект може да бъде използван конструктивно – като инструмент за автоматизирана, насочваща обратна връзка, която подпомага формиращото оценяване, без да подменя експертната роля на преподавателя.

Изследването е направено в рамките на научен проект „Разговорната реч в дигитален формат – от архив към научен ресурс“ ***(ФСД-31-354-14/28.04.2026 г.)****.*

ЛИТЕРАТУРА

1. Тодоранова, Л. (2025). Изкуственият интелект – инструмент за обратна връзка в обучението. В: *Трета национална научно-практическа конференция „Дигитална трансформация на образованието – проблеми и решения“. Сборник доклади*. Русе: Академично издателство на Русенския университет, 533–537.

2. Byanov, I. (2025). Прилагането на стратегическите планове в рамките на Общата селскостопанска политика: иновация или продължение на досегашната практика In: The Economy of the 21st Century: Economic Innovations and Sustainable Growth. Sofia: NBU, 32–44, ISBN: 978-619-233-368-3.

3. Clark, E., T. August, S. Serrano, N. Haduong, S. Gururangan & N. A. Smith. (2021). All That’s ‘Human’ Is Not Gold: Evaluating Human Evaluation of Generated Text. In: *Proceedings of the 59th Annual Meeting of the Association for Computational*

Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 7282–7296*,* DOI:10.18653/v1/2021.acl-long.565

4. Gehrmann, S., H. Strobelt & A. Rush. (2019). GLTR: Statistical Detection and Visualization of Generated Text. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 111–116, DOI:10.18653/v1/P19-3019

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

6. Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S. et al. (2023). Testing of detection tools for AI-generated text. *Int J Educ Integr,* 19, (26) DOI:10.1007/s40979-023-00146-z