

Człowiek o AI (8)

* Student, Uniwersytet Warszawski

Mintz, S., *Apparently, I'm a Bot*, Inside Higher Ed, kwiecień 2025.

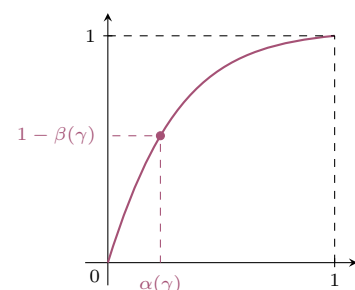
Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., Feizi, S., *Can AI-Generated Text be Reliably Detected?*, arXiv:2303.11156v4, preprint, 2025. Kod źródłowy eksperymentów dostępny pod adresem: github.com/vinusankars/Reliability-of-AI-text-detectors.

Kirchenbauer, J. i in., *A Watermark for Large Language Models*, 2023.

Li, Y. i in., *MAGE: Machine-Generated Text Detection in the Wild*, 2024.

Krishna, K. i in., *Paraphrasing Evades Detectors of AI-Generated Text, but Retrieval is an Effective Defense*, 2023.

Mitchell, E. i in., *DetectGPT: Zero-Shot Machine-Generated Text Detection Using Probability Curvature*, 2023.



Przykładowa krzywa ROC

Nie ma tekstu zbyt ludzkiego...

Jakub RYŁOW*

Steven Mintz, profesor historii na Uniwersytecie Tekszańskim w Austin, napisał w swojej karierze ponad osiemset esejów. Kiedy pewnego dnia wrzucił jeden z nich do popularnego detektora AI, otrzymał wynik: **100% wygenerowane przez sztuczną inteligencję**. Tekst opisywał spektakl muzyczny, który Mintz widział osobiście, w sali z pięćdziesięcioma widzami, bez żadnego nagrania ani śladu w internecie. Detektor nie przejął się tym szczegółem. To samo narzędzie oceniło przemówienie gettynburskie Lincolna jako stuprocentowo maszynowe, a esej Orwella *Shooting an Elephant* z 1936 roku jako podejrzany w ponad połowie. Jest w tym pewna poetycka sprawiedliwość: dzieło człowieka będącego twórcą opisu świata, w którym myślozbrodnię wykrywa się na podstawie stylu wypowiedzi, pada ofiarą algorytmu wykrywającego myślozbrodnię na podstawie stylu wypowiedzi. Orwell napisał *Rok 1984* jako ostrzeżenie. Nikt nie spodziewał się, że będzie też instrukcją obsługi detektora plagiatów. Atmosfera podejrzliwości wobec dobrze napisanego tekstu jest bowiem jak najbardziej orwellowska: im staranniej piszesz, tym bardziej jesteś podejrzany.

Można by uznać te przypadki za błędy implementacji konkretnych narzędzi. Jednak w 2025 roku Sadasivan, Kumar, Balasubramanian, Wang i Feizi opublikowali artykuł *Can AI-Generated Text be Reliably Detected?*, w którym wykazali, że problem nie leży w implementacji. Problem leży w naturze zadania, co wyjaśnimy poniżej.

Rodzaje detektorów

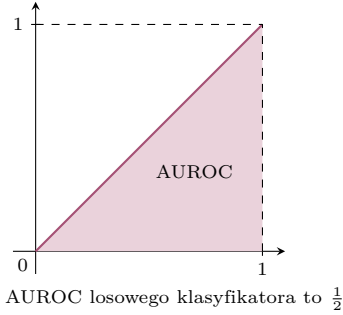
Istnieją różne sposoby na wykrywanie tekstów wygenerowanych przez AI. Mamy na przykład detektory oparte na tzw. *znakach wodnych*. Niektóre modele celowo generują tekst z ukrytym wzorcem statystycznym, preferując tokeny z tzw. zielonej listy, by możliwe było wykrycie źródła pochodzenia tekstu metodami analizy statystycznej. Druga klasa to detektory oparte na sieciach neuronowych, trenowane binarnie na parach tekstów ludzkich i maszynowych. Trzecia to detektory *oparte na wyszukiwaniu*: dla danego tekstu sprawdzają, czy nie ma semantycznie doń podobnego w bazie wielu tekstów wygenerowanych przez konkretny model. Oczywiście są również inne pomysły. Na przykład narzędzie DetectGPT opiera się na obserwacji, że (w dużym uproszczeniu) jeśli model maszynowy lekko zmodyfikuje tekst wygenerowany przez maszynę, to efekt nie powinien wyglądać na jeszcze bardziej maszynowy.

Jak oceniane są detektory?

Detektory nie tylko oceniają, ale i same są oceniane – jak to się robi? Możemy w (nie tak znowu dużym) uproszczeniu przyjąć, że każdy detektor to pewna funkcja D , która dowolnemu tekstowi s przypisuje liczbę $D(s) \in [0, 1]$. Liczba ta jest siłą przekonania detektora o tym, że tekst s został wygenerowany przez AI. Ostatecznie jednak badacz chce być w stanie rozróżniać między tekstami sztucznymi a wytworzonymi przez człowieka. W tym celu ustala pewien próg $\gamma \in [0, 1]$ i wszystkie teksty s , dla których $D(s) \geq \gamma$ uznaje za sztuczne. Oczywiście zdaje on sobie sprawę z tego, że zapewne popełnia błędy. Mogą one być dwojakiego rodzaju: może błędnie uznawać teksty człowieka za sztuczne (jak w opisanych na wstępie przykładach) oraz odwrotnie. Prawdopodobieństwo popełnienia pierwszego typu błędu to $\alpha(\gamma) := \mathbb{P}_H(D(S) \geq \gamma)$, tzn. zakładamy, że S jest losowym słowem wytworzonym przez człowieka. Analogicznie, szansa na drugi błąd to $\beta(\gamma) := \mathbb{P}_{AI}(D(S) < \gamma)$.

Nietrudno zauważyć, że funkcja $\alpha(\gamma)$ jest malejąca, zaś funkcja $\beta(\gamma)$ jest rosnąca. Punkty $(\alpha(\gamma), 1 - \beta(\gamma))$ dla wszystkich $\gamma \in [0, 1]$ tworzą zatem pewną „rosnącą” krzywą, łączącą punkt $(0, 0)$ z punktem $(1, 1)$. Określa się ją skrótem ROC (ang. *Receiver Operator Characteristic*; nazwa sięga II Wojny Światowej, za jej pomocą oceniano skuteczność radarów). Od dobrego detektora oczekujemy, że można do niego dobrać próg γ dający możliwie niskie wartości $\alpha(\gamma)$ i $\beta(\gamma)$,

tn. że jego krzywa ROC przechodzi możliwie blisko (w jakimś sensie) punktu $(0, 1)$. Wobec monotoniczności ROC sensownym sposobem pomiaru tej bliskości jest pole powierzchni obszaru przez nią ograniczonego. Jest to klasyczny sposób pomiaru jakości detektora (nie tylko tekstów), oznaczany jako AUROC(D) (*Area Under ROC*) lub, trochę krócej, AUC(D).



Zobaczmy, jak działają te pojęcia na przykładzie dość żalostnego detektora, który nawet nie patrzy na dostarczony mu tekst s , tylko daje losową liczbę z przedziału $[0, 1]$. Wówczas $\beta(\gamma) = 1 - \alpha(\gamma) = \gamma$, więc krzywa ROC to wykres funkcji $y = x$ na $[0, 1]$, a także AUROC = $\frac{1}{2}$. Jest to również rozsądny punkt odniesienia – każdy sensowny detektor powinien mieć AUROC większe od $\frac{1}{2}$.

AUROC a odległość między rozkładami

Nie zaskoczy nikogo stwierdzenie, że im bardziej podobne do ludzkich teksty generuje sztuczna inteligencja, tym trudniejsze zadanie stoi przed detektorami. Można jednak mieć nadzieję na to, że jeśli tylko jest między tymi źródłami tekstów jakakolwiek „statystyczna” różnica, to odpowiednio „silny” detektor będzie sobie rozsądnie radził z odróżnianiem jednych od drugich.

Okazuje się, że nie jest tak dobrze. Aby się o tym przekonać, wprowadźmy następujący sposób mierzenia odległości między losowością tekstów ludzkich i maszynowych. Jest to tzw. *odległość całkowitego wahania* między dwoma rozkładami prawdopodobieństwa. Niech, tak jak wcześniej, AI oznacza losowość (formalnie: rozkład prawdopodobieństwa) wszystkich możliwych sekwencji tokenów generowanych przez model językowy, a H – analogiczną charakterystykę tekstów ludzkich. Wtedy odległość całkowitego wahania między nimi określona jest wzorem

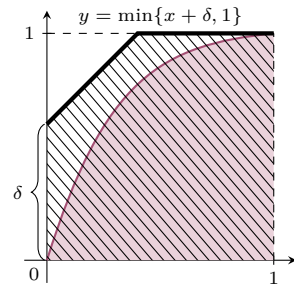
$$TV(H, AI) = \sup_{A \subseteq \Omega} |\mathbb{P}_{AI}(S \in A) - \mathbb{P}_H(S \in A)|,$$

gdzie Ω jest przestrzenią wszystkich możliwych tekstów. Intuicja jest raczej prosta – pytamy o maksymalną, bezwzględną różnicę między prawdopodobieństwami wygenerowania tekstu należącego do zbioru A dla wszystkich możliwych wyborów A .

Zauważmy, że

$$|(1 - \beta(\gamma)) - \alpha(\gamma)| = |\mathbb{P}_{AI}(D(s) \geq \gamma) - \mathbb{P}_H(D(s) \geq \gamma)| \leq TV(H, AI).$$

Oznaczając prawą stronę powyższej nierówności przez δ , otrzymujemy, że krzywa ROC jest ograniczona z góry przez wykres funkcji $y = x + \delta$, a nawet $y = \min\{x + \delta, 1\}$. Miare AUROC jakości detektora możemy zatem oszacować z góry przez pole pod wykresem funkcji $y = \min\{x + \delta, 1\}$, czyli $\frac{1}{2} + \delta - \frac{\delta^2}{2}$. Morał: przy odpowiednio małych wartościach odległości całkowitego wahania *każdy* detektor staje się niewiele bardziej użyteczny niż losowe zgadywanie.



AUROC (oznaczone kolorem) szacuje się z góry przez pole zakresowanego obszaru, czyli $\frac{1}{2} + \delta - \frac{\delta^2}{2}$, gdzie $\delta = TV(H, AI)$

Atak przez parafrazowanie: jak AI ukrywa AI

Powyższa obserwacja dotyczy detektorów w świecie bez atakujących. Sadasivan i współautorzy pokazują, że w świecie z atakującymi sytuacja jest jeszcze gorsza. Wystarczy *rekurencyjne parafrazowanie*: tekst S generowany przez model L zastępujemy przez $P(S)$, następnie $P(P(S))$ i tak dalej, gdzie P jest dowolnym modelem parafrazującym.

Po pięciu rundach z modelem DIPPER (11 miliardów parametrów) wskaźnik wykrywalności znaków wodnych spada z 99,3% do 9,7%. Już dwie rundy wystarczają, by zejść poniżej 50% wykrywalności. AUROC detektora DetectGPT spada po parafrazie z 0,965 do 0,25 – poniżej wartości odpowiadającej losowemu zgadywaniu. Jakość parafraz prawie nie cierpi: 89% oceniono w badaniu MTurk jako gramatycznie doskonałe.

Spójność z powyższymi obserwacjami jest pełna: parafrazowanie zbliża rozkład przetworzonego tekstu do rozkładu ludzkiego, zmniejsza odległość całkowitego wahania, a wraz z nią skuteczność detektora.

Touvron, H. i in., *Llama 2: Open Foundation and Fine-Tuned Chat Models*, arXiv preprint, 2023. LLaMA-2-7B-Chat pełnił rolę „parafrazera”. Paradoksalnie: AI ukrywa AI.

Zhang, S. i in., *OPT: Open Pre-Trained Transformer Language Models*, arXiv preprint, 2022. Modele OPT-1.3B i OPT-13B były celami ataków.

Rajpurkar, P. i in., *SQuAD: 100 000+ Questions for Machine Comprehension of Text*, 2016. Benchmark SQuAD-v2 potwierdził zachowanie treści po parafrazie.

Gu, C. i in., *On the Learnability of Watermarks for Language Models*, arXiv preprint, 2024.

Quach, K., *Professor Freezes Student Grades After ChatGPT Claimed AI Wrote Their Papers*, The Register, 2023. Studenci pisali. Profesor zamroził oceny. Nierówność $AUROC \leq \frac{1}{2} + \delta - \frac{\delta^2}{2}$ kiwała głową.

Kosmyna, N. i in., *Your Brain on ChatGPT: Accumulation of Cognitive Debt When Using an AI Assistant for Essay Writing Task*, 2025. Badanie wstępne, przed pełną recenzją; autorzy zalecają ostrożność interpretacyjną.

Gerlich, M., *AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking*, 2025. Badanie na 666 uczestnikach potwierdziło mediującą rolę tego zjawiska między używaniem AI a osłabieniem myślenia krytycznego.

Salatino, A. i in., *Influence of AI Behavior on Human Moral Decisions, Agency, and Responsibility*, 2025. Bierne poleganie na AI osłabia nie tylko pamięć, lecz także poczucie sprawczości i odpowiedzialności moralnej.

Liang, W. i in., *GPT Detectors are Biased Against Non-Native English Writers*, 2023. Detektory karzą też obcokrajowców piszących po angielsku – za pisanie z akcentem.

Podobny los spotkał dwukropek, wyliczenia i zdania wielokrotnie złożone. Interpunkcja jako dowód winy.

Zob. Sourati, Z., Ziabari, A.S., Dehghani, M., *The Homogenizing Effect of Large Language Models on Human Expression and Thought, Trends in Cognitive Sciences*, marzec 2026; oraz popularnonaukowe omówienie: Chayka, K., *A.I. Is Homogenizing Our Thoughts*, *The New Yorker*, 25 czerwca 2025.

Fowler, G. A., *We Tested a New ChatGPT-Detector for Teachers. It Flagged an Innocent Student*, *The Washington Post*, 2023; Al-Sibai, N., *AI Plagiarism Detection Software Keeps Falsely Accusing Students of Cheating*, *Futurism*, 2023.

Kiedy detektor atakuje niewinnych

Artykuł bada też kierunek odwrotny: jak sprawić, by tekst *ludzki* wyglądał jak tekst wyprodukowany przez AI. W przypadku metod opartych na znakach wodnych wystarczy wielokrotnie zapytać model o następnik danego słowa, by statystycznie odtworzyć zieloną listę, która z założenia powinna być prywatna. Dysponując tą wiedzą, napastnik komponuje zdania bogate w tokeny z zielonej listy i detektor uznaje go za znakowany model. AUROC detektora wodnego spada w eksperymencie z 0,998 do 0,013 – narzędzie klasyfikuje teksty ludzkie jako maszynowe z taką samą pewnością, z jaką wcześniej klasyfikowało poprawnie.

Scenariusz z detektorem opartym na wyszukiwaniu jest jeszcze prostszy i etycznie bardziej niepokojący. Nauczyciel prosi model o sparafrazowanie eseju studenta. Model zapisuje parafrazę w swojej bazie. Detektor porównuje oryginał studenta z parafrazą i ogłasza plagiat. Student pisał sam. Matematycznie trywialne. Etycznie nieakceptowalne.

Mózg na zwolnieniu: po co w ogóle pisać samemu

Do tej pory mówiliśmy o tym, co AI robi z detektorami. Warto zapytać, co robi z piszącymi. Wyniki są równie niepokojące. Zespół MIT Media Lab pod kierownictwem Natalii Kosmyny wyposażył 54 studentów z obszaru Bostonu w czepki EEG i przez cztery miesiące obserwował aktywność ich mózgów podczas pisania esejów. Aktywność mierzona wskaźnikiem Dynamic Directed Transfer Function (dDTF) – przepływem informacji między obszarami mózgu odpowiedzialnymi za funkcje wykonawcze, przetwarzanie semantyczne i regulację uwagi – była niższa o 55% u studentów piszących z pomocą ChatGPT niż u piszących bez żadnych narzędzi. 83% studentów z grupy ChatGPT nie potrafiło po zakończeniu badania przypomnieć sobie kluczowych twierdzeń ze swoich własnych esejów. Efekty utrzymywały się nawet po „odstawieniu” modelu: studenci, którzy zaczęli od ChatGPT, a potem pisali samodzielnie, wykazywali niższą aktywność neuronową niż ci, którzy od początku pracowali bez wspomagania.

Mechanizm leżący u podstaw tego zjawiska to *cognitive offloading* – przenoszenie wysiłku poznawczego na narzędzie zewnętrzne. Skumulowany efekt powtarzanego delegowania wysiłku intelektualnego badacze nazywają *dlugim poznawczym* – oprocentowanym zaciąganiem pożyczki od własnego mózgu. Mózg rządzi się zasadą *use it or lose it*: obwody regularnie nieużywane ulegają osłabieniu.

Za co naprawdę karzą detektory

Wróćmy do Mintza, Lincolna i Orwella. Detektory nie mierzą proveniencji. Mierzą wzorce powierzchniowe: formalność stylu, regularność składni, brak potknięć, przemyślane przejścia między akapitami. Dobra proza ma dokładnie te cechy. W świecie detektorów nie ma tekstu zbyt ludzkiego – jest tylko tekst na tyle niestaranny, by nie wzbudzić podejrzeń.

Do podejrzanych sygnałów należy też pauza – znak interpunkcyjny obecny w prozie od Stendhala po Miłosza, stosowany wszędzie tam, gdzie przecinek jest za słaby, a nawias zbyt urzędowy. Detektory nauczyły się, że AI lubi pauzy, i wyciągnęły z tego wniosek odwrotny do zamierzonego: że każdy, kto ich używa, jest podejrzany.

Logiczną konsekwencją jest homogenizacja: pisz jak wszyscy, myśl jak wszyscy, nie wyróżniaj się nawet minimalnie ponad popositą średnią – bo niestandardowość jest dziś dowodem winy. Efekt jest dość groteskowy: pewien serwis oferuje za opłatą uszlachetnianie tekstu w kierunku usterkowości – przez celowe wstawienie błędów gramatycznych i zastąpienie formalnych łączników czymś mniej starannym. Żeby brzmieć jak człowiek, trzeba nauczyć się pisać gorzej.

Zestawienie obu wątków ujawnia paradoks strukturalny. Studenci używający AI piszą szybciej, ale pamiętają mniej i tracą poczucie autorstwa. Studenci

piszący bez AI, lecz dobrze, są oskarżani o użycie AI. Obserwacja Sadasivana i współautorów wyjaśnia, dlaczego ten paradoks jest nieusuwalny: gdy $TV(M, H)$ jest małe, detektor nie może odróżnić obu grup – i nie odróżnia.

Warto jednak odnotować, że entuzjazm, z jakim niektórzy rozsyłają dowody bezskuteczności detektorów, jest sam w sobie zastanawiający. *The lady doth protest too much, methinks...*

Szekspir, W., *Hamlet*, akt III, scena 2.

Konkluzja

Detektory mogą być użytecznym narzędziem pomocniczym pod warunkiem, że nikt nie traktuje ich wyniku jako dowodu. Pewnym paradoksem jest fakt, że konsekwencją postępującego rozwoju technologicznego (który w teorii miał gwarantować wiarygodność i weryfikować wiele kwestii niemierzalnych, takich jak zaufanie) jest rozmycie tej pewności i precyzji w wielu aspektach, pozostawiając dużo miejsca na domysły, tutaj akurat w kwestii przypisania autorstwa do wybranego dzieła. Jednocześnie otwiera to drogę na zupełnie nieścisłe metody weryfikacji: intuicję, przecucie, wrażenie, czy dany tekst powstał samodzielnie, czy nie.

Bot – program komputerowy, który nie jest człowiekiem, lecz próbuje imitować jego działania, np. tworzyć teksty czy prowadzić rozmowy.

Das, M. R., *Turn-It-In: AI Fails Students for Not Using AI*, Firstpost, 2023; Weiss, M., *Deepfake Bot Submissions to Federal Public Comment Websites Cannot Be Distinguished from Human Submissions*, Technology Science, 2019.

Profesor Mintz jest botem. Lincoln jest botem. Orwell – autor ostrzegający przed światem, w którym myśl jest przestępstwem – jest botem. ChatGPT przechodzi test człowieczeństwa z wynikiem lepszym niż wszyscy trzej razem wzięci. Studenci, którzy oddali pisanie modelowi, nie pamiętają własnych esejów. Studenci, którzy pisali samodzielnie i pisali dobrze, zostali oskarżeni o plagiat. Nie ma tekstu wystarczająco ludzkiego, są tylko teksty za mało dokładnie przetestowane na użycie AI.

Tekst powstał bez użycia AI.

O badaniu i pożytkach z siarczku molibdenu

*Wydział Chemii oraz Centrum Nauk Biologiczno-Chemicznych, Uniwersytet Warszawski
rszoszkiewicz@chem.uw.edu.pl

Robert SZOSZKIEWICZ*

Fizykochemia powierzchni, która wyłoniła się wskutek używania teorii, metod i narzędzi badawczych z zakresu zarówno fizyki, chemii, jak i inżynierii materiałowej, bardzo pręźnie rozwijała się już w latach 80. i 90. ubiegłego wieku. Obecnie, wraz pojawieniem się tzw. materiałów dwuwymiarowych (2D) oraz niesamowitego rozwoju rozmaitych metod mikroskopowych, przeżywa kolejny renesans.

W terminologii naukowej taką operację złuszczenia określamy mianem *eksfoliacji mechanicznej materiałów objętościowych*. Po jej dokładniejszy opis można sięgnąć do artykułu *O tajemnicach eksfoliacji, czyli co łączy grafen z odklejaniem etykiet z cenami z Δ_{15}^3* .

Materiały dwuwymiarowe to pewna klasa materiałów krystalicznych, których właściwości chemiczne i fizyczne bardzo silnie zależą od grubości. Z reguły są to niemalże atomowo cienkie warstwy. W najprostszy sposób można je otrzymać poprzez złuszczenie mikroskopowych kryształów. Taka operacja jest możliwa dlatego, że atomy w poszczególnych warstwach materiałów dwuwymiarowych połączone są z sobą silnymi wiązaniami kowalencyjnymi, zaś poszczególne warstwy łączą się z sobą słabymi oddziaływaniami o naturze elektrostatycznej, określanymi mianem oddziaływań van der Waalsa. Ze względu na skład atomowy i właściwości chemiczne wyróżnić można wiele klas takich materiałów, na przykład: grafen i jego pochodne, azotek boru i jego pochodne, dichalkogenki metali przejściowych, perowskity i *MXeny*.

O grafenie można było przeczytać w *Delcie*, w artykule *Grafen i spółka*, Δ_{21}^5 .

Rewolucja w świecie materiałów 2D została zapoczątkowana przez otrzymanie w 2004 roku *grafenu*, czyli pojedynczej warstwy atomów węgla ułożonych w jednej płaszczyźnie, w tzw. sieci heksagonalnej, a więc strukturze przypominającej plaster miodu, tyle że atomowo cenniejszy. Okazało się, że ma on – przewidziane przez wcześniejsze rozważania teoretyczne – niezwykle własności, jak na przykład niemalże bezstratne przewodnictwo elektronowe.