# Czy ten artykuł napisała sztuczna inteligencja? Studium przypadku „wykrywania plagiatu”

Anna Małgorzata Kamińska
ORCID: 0000-0002-8935-9735
*Instytut Nauk o Kulturze*
*Wydział Humanistyczny*
*Uniwersytet Śląski w Katowicach*

**Abstrakt**

**Cel/Teza:** Artykuł ocenia użycie detektorów treści generowanych przez GenSI w kontekście weryfikacji uczciwości akademickiej. Główna teza głosi, że narzędzia te nie potrafią odróżnić tekstu wygenerowanego przez SI od treści opracowanych przez człowieka poddanych jedynie redakcji stylistycznej przez duże modele językowe, przez co wysokie wyniki detekcji nie stanowią wiarygodnego dowodu plagiatu.
**Koncepcja/Metody badań:** Zastosowano studium przypadku. Stworzono biografię fikcyjnej postaci autorstwa człowieka, którą następnie zredagowano przy użyciu ChatGPT i Claude z precyzyjnymi poleceniami ograniczającymi SI do roli redakcyjnej. Wszystkie wersje przeanalizowano detektorem ZeroGPT.
**Wyniki i wnioski:** Tekst oryginalny uzyskał wynik 0% prawdopodobieństwa bycia wygenerowanym przez SI, podczas gdy wersje zredagowane przez ChatGPT i Claude osiągnęły odpowiednio 96,7% i 100%. Detektory identyfikują cechy stylistyczne charakterystyczne dla dobrze zredagowanych tekstów, a nie rzeczywiste pochodzenie wkładu intelektualnego.
**Zastosowania praktyczne:** Wyniki wskazują na konieczność rewizji polityk akademickich opartych na detektorach SI. Instytucje powinny promować transparentne zasady wykorzystania SI jako narzędzia wspierającego, zamiast polegać na zawodnych technologiach prowadzących do niesprawiedliwych ocen.
**Oryginalność/Wartość poznawcza:** Praca dostarcza empirycznych dowodów na brak wiarygodności detektorów SI w kontekście wykrywania plagiatów, wyznaczając jasną granicę etyczną między wykorzystaniem SI jako narzędzia a nieuczciwością akademicką oraz promując świadome podejście do wdrażania nowych technologii w nauce.

**Słowa kluczowe**

Detekcja treści SI. Duże modele językowe (DMJ). Etyka w nauce. Generatywna sztuczna inteligencja. GenSI. Komunikacja naukowa. Narzędzia redakcyjne. Plagiat. Studium przypadku.

*Tekst wpłynął do Redakcji 28 września 2025 r.*

## 1. Wstęp

„Czy ten artykuł napisała sztuczna inteligencja?" – takie i podobne pytania, czytając treści akademickie, zadaje sobie obecnie coraz więcej osób: od recenzentów i redaktorów czasopism naukowych, po wykładowców sprawdzających prace studentów. Wynika to z faktu, że pojawienie się i gwałtowna popularyzacja generatywnej sztucznej inteligencji (GenSI), w szczególności dużych modeli językowych (DMJ), stanowi jeden z najbardziej przełomowych momentów w najnowszej historii rozwoju technologii. Ich wpływ na proces tworzenia, przetwarzania i konsumpcji treści jest niezaprzeczalny i dotyka niemal każdej dziedziny życia, w tym – w sposób szczególny i wielowymiarowy – środowiska akademickiego (Tillmanns *et al.*, 2025; Yusuf *et al.*, 2024). DMJ, takie jak modele z rodziny GPT firmy OpenAI czy Claude firmy Anthropic, przestały być technologiczną ciekawostką, a stały się zaawansowanymi narzędziami, które mogą służyć jako asystenci w badaniach, programowaniu, analizie danych, a przede wszystkim – w pracy z tekstem.

Ta technologiczna transformacja stała się jednak źródłem fundamentalnego konfliktu. Z jednej strony, naukowcy i autorzy otrzymali bezprecedensowe wsparcie w procesie pisarskim, umożliwiające im poprawę klarowności swoich prac, a co za tym idzie ułatwienie zrozumienia treści i obniżenie prawdopodobieństwa błędnych ich interpretacji przez potencjalnych czytelników – co ma pozytywny wpływ na proces upowszechniania wiedzy naukowej, a tym samym służy dobru środowiska akademickiego. Z drugiej jednak strony, łatwość, z jaką DMJ potrafią generować obszerne i spójne językowo teksty na dowolny temat, wywołała uzasadnione obawy o nieuczciwość akademicką, integralność badań naukowych, a nawet nową formę plagiatu (Francis *et al.*, 2025). W odpowiedzi na tak artykułowane zagrożenia, pojawiły się liczne programy, których celem jest detekcja treści rzekomo wygenerowanych przez GenSI. Doprowadziło to do zjawiska, które można określić mianem „paniki moralnej", w której używanie narzędzia wykrywającego tekst redagowany przez DMJ wydaje się prostym i skutecznym remedium na złożony problem etyczny.

Niniejszy artykuł stawia jednak tezę, że bezkrytyczne poleganie na takich założeniach i rozwiązaniach jest nie tylko nieskuteczne, ale wręcz szkodliwe dla procesu komunikacji naukowej. W pracy postawiono następujące pytania badawcze:

(1) Czy tekst, którego autorem merytorycznym w 100% jest człowiek, może zostać błędnie zakwalifikowany jako dzieło SI wyłącznie po przeprowadzeniu na nim korekty językowej za pomocą DMJ?
(2) Gdzie przebiega granica między etycznym wykorzystaniem SI jako zaawansowanego narzędzia redakcyjnego a nieuczciwością akademicką?
(3) Jakie są potencjalne konsekwencje polegania na detektorach SI dla studentów, nauczycieli, badaczy, recenzentów, redaktorów czasopism i całego ekosystemu naukowego?

Argumentem przewodnim artykułu jest stwierdzenie, że wykorzystanie DMJ w roli „cyfrowego redaktora" jest naturalną i logiczną ewolucją wielu od dawna już używanych narzędzi wspomagających pisanie i samo w sobie nie stanowi plagiatu. Jednocześnie, obecne systemy detekcji są z natury niezdolne do rozróżnienia między treścią wygenerowaną (gdzie SI jest autorem merytorycznym) a zredagowaną (gdzie SI jest jedynie wykorzystywane w roli asystenta redakcyjnego), co stwarza poważne ryzyko fałszywych oskarżeń i może hamować rozwój dobrych praktyk w komunikacji naukowej. W celu udowodnienia tej tezy, w dalszej części pracy przedstawiono studium przypadku, które w kontrolowanych warunkach ilustruje opisywany problem.

## 2. Tło teoretyczne i przegląd literatury

Technologiczne wspomaganie procesu pisania nie jest zjawiskiem nowym. Jego historia sięga pierwszych edytorów tekstu wyposażonych w proste mechanizmy sprawdzania pisowni. Kolejnym krokiem było wprowadzenie korektorów gramatycznych, które potrafiły analizować podstawowe struktury zdaniowe (Oakman, 1994). Przełomem stały się narzędzia takie jak Grammarly, które dzięki analizie dużych zbiorów danych oferowały sugestie stylistyczne, synonimy i propozycje poprawy płynności tekstu (Barrot, 2022). Z tego punktu widzenia, DMJ są kolejnym, choć znacznie bardziej zaawansowanym, ogniwem w tym samym łańcuchu ewolucyjnym, patrząc zarówno z perspektywy funkcjonalnej, jak i technologicznej. Ich zdolność do identyfikacji kontekstów, niuansów językowych i skomplikowanych struktur narracyjnych sprawia, że mogą one pełnić rolę zaawansowanego asystenta w redakcji tekstu, a nie tylko pasywnego korektora. Traktowanie ich jako całkowicie nowej, niedopuszczalnej, kategorii narzędzi ignoruje ten historyczny kontekst rozwoju technologicznego. Głosy optujące za całkowitym wykluczeniem adopcji tej technologii, choć mające podstawy – zwłaszcza u osób, które jej nie rozumieją i w wyniku tego obawiają się jej – stoją w sprzeczności z ogólną ideą postępu. Zjawisko to jest w zasadzie powtarzalne, a ludzkość doświadczała go przy okazji pojawienia się każdej z przełomowych innowacji. Zwykle jednak, po fazie sceptycyzmu i strachu, zaczyna się stawiać pytania nie o to: czy, ale o to: jak wykorzystywać daną technologię dla dobra cywilizacji i zrównoważonego rozwoju ludzkości.

W świetle narastających kontrowersji w obszarze zastosowania narzędzi GenSI w praktykach edukacyjnych i badawczych, konieczne staje się precyzyjne określenie, gdzie kończy się rola GenSI jako dopuszczalnego narzędzia, a zaczyna się jego zastosowanie związane z nieuczciwością akademicką, co wymaga ponownego przemyślenia klasycznej definicji plagiatu w erze DMJ (Luo, 2024).

Klasyczna definicja plagiatu, głęboko zakorzeniona w etyce akademickiej, odnosi się do przywłaszczenia cudzych idei lub danych i przedstawienia ich jako

własnych, bez odpowiedniego odniesienia do źródła (Helgesson and Eriksson, 2015). Kluczowym elementem jest tu zatem pochodzenie wkładu intelektualnego – merytorycznej treści pracy. W przypadku, gdy autor wykorzystuje DMJ do poprawy interpunkcji, przeformułowania niejasnego zdania czy znalezienia lepszego synonimu, nie dochodzi do przywłaszczenia cudzych idei. Autor wciąż pozostaje w pełni odpowiedzialny za każdy fakt, argument i tezę zawartą w tekście. Źródłem wiedzy jest autor, a model językowy staje się jedynie narzędziem, które tę wiedzę pomaga ubrać w bardziej przystępną formę językową. Rola DMJ jest w tym kontekście funkcjonalnie tożsama z rolą ludzkiego redaktora lub kolegi poproszonego o przeczytanie i zaopiniowanie czytelności i jasności tekstu.

Niestety, pomimo etycznej czystości takiego wykorzystania GenSI, środowisko akademickie szybko zareagowało na to zjawisko, koncentrując się nie na intencji autora, lecz na technologicznej sygnaturze tekstu, co bezpośrednio doprowadziło do gwałtownego rozwoju internetowych usług oferujących możliwość wykrywania tekstu generowanego przez SI. Liczne badania naukowe i raporty techniczne wskazują jednak na fundamentalne problemy z ich skutecznością i wiarygodnością. Wykazano, że narzędzia te mają tendencję do generowania dużej liczby wyników fałszywie pozytywnych (ang. *false positives*), oznaczając teksty napisane przez ludzi jako dzieło SI. Problem ten jest szczególnie dotkliwy w przypadku tekstów, które z natury są ustrukturyzowane i posługują się specjalistycznym, powtarzalnym językiem, jak ma to miejsce w tekstach naukowych, prawniczych czy medycznych.

Co więcej, detektory często identyfikują jako „sztuczną" cechę, którą określa się jako niską perpleksję – czyli wysoki stopień przewidywalności kolejnych słów w zdaniu. Jest to cecha charakterystyczna dla tekstów generowanych przez DMJ, ale jednocześnie to pożądana właściwość klarownego i dobrze napisanego tekstu, ponieważ w nauce klarowność i jednoznaczność są niezbędne do precyzyjnego przekazywania wyników, replikacji badań i budowania wiedzy. Oznacza to, że tekst naukowy powinien być przewidywalny składniowo (czyli mieć niską perpleksję), aby odbiorca mógł skupić się na merytoryce, a nie na interpretowaniu zawiłości językowych. W rezultacie, im bardziej autor stara się pisać płynnie, spójnie i zrozumiale, tym większe ryzyko, że jego tekst zostanie oznaczony przez detektor. Tworzy to paradoksalną sytuację, w której narzędzia mające stać na straży jakości, mogą w istocie karać za dążenie do niej.

Z drugiej strony barykady, bezkrytyczne traktowanie wyników detektorów SI jako dowodu nieuczciwości natychmiast wywołało kontrreakcję w postaci rozwoju narzędzi mających na celu „humanizowanie" tekstu. Narzędzia te działają na zasadzie celowego zaburzania płynności tekstu generowanego przez GenSI. Ich zadaniem jest wprowadzenie do spójnych i logicznie przewidywalnych fraz elementów o wysokiej perpleksji (symulowanie błędów, niespójności lub nadmiernie skomplikowanego słownictwa), co ma na celu obniżenie wskaźnika wykrywalności

przez detektory. Proces ten, w pełni automatyczny i masowy, pozwala na błyskawiczne „psucie" treści generowanych przez sztuczną inteligencję w celu symulowania ludzkich niedoskonałości redakcyjnych. Istnienie tej kategorii narzędzi dowodzi, że w obecnej sytuacji technologicznej nie ma już problemu z tym, aby treści generowane automatycznie były również automatycznie i masowo maskowane przed detektorami, co dodatkowo obnaża nieprzydatność tych ostatnich w roli arbitra etyki akademickiej.

## 3. Metodyka studium przypadku

W celu empirycznej weryfikacji postawionej tezy badawczej przeprowadzono eksperymentalne studium przypadku. Taki wybór metody badawczej jest uzasadniony koniecznością osiągnięcia wiarygodności i precyzji wynikającej z kontrolowanego przebiegu badania, a jednocześnie zapewnia pogłębione, jakościowe spojrzenie typowe dla takiego studium. Zdolność do szczegółowej analizy badanego zjawiska w ściśle kontrolowanych warunkach umożliwia jednoznaczną ilustrację i precyzyjną egzemplifikację problematyki zawodności detektorów SI.

Na potrzeby badania opracowano „surowy" tekst biograficzny fikcyjnej postaci – polskiego jazzmana, Bronisława „Borsuka" Wierzbickiego, rzekomo aktywnego w okresie powojennej odwilży. Można jednak założyć, że biogram ten został napisany przez hipotetycznego historyka na podstawie trudno dostępnych, niezdigitalizowanych źródeł, a co za tym idzie – nieobecnych również w przestrzeniach internetowych (archiwalia, wywiady, prasa z epoki). Niewątpliwie więc treść merytoryczna tego biogramu, będąca syntezą specjalistycznej wiedzy, odzwierciedla wkład intelektualny hipotetycznego autora. Autor ten, tworząc roboczą wersję notatki, traktował ją bardziej jako szkic badawczy, nie przykładając jeszcze na tym etapie należytej wagi do płynności użytych opisów, a nawet dopuszczając przez roztargnienie błędy ortograficzne czy stylistyczne. Wybór fikcyjnej postaci Bronisława „Borsuka" Wierzbickiego był zamierzonym zabiegiem metodologicznym, który miał na celu wykazać, że modele mogą wspierać redakcję tekstu dotyczącego nieznanej im tematyki. Co istotne, imię i nazwisko postaci zostały celowo zdublowane z tożsamością innego, dobrze znanego w przestrzeni internetowej (a tym samym potencjalnie obecnego w danych treningowych modeli), bohatera fikcyjnego – postaci z serialu filmowego *Ojciec Mateusz*. Taki zabieg miał na celu przeprowadzenie najbardziej przekonującego testu: wykazanie, że poprzez precyzyjne polecenia można całkowicie pominąć wiedzę DMJ w danej dziedzinie, wykorzystując model jedynie w roli pomocnego asystenta redakcyjnego, ze szczególnym naciskiem na jego umiejętności lingwistyczne. Stanowiło to jednocześnie dodatkowy sprawdzian zdolności modeli do ścisłego podążania za poleceniami i nieulegania tendencji do halucynacji lub włączania do tekstu zewnętrznych informacji.

Poniżej zaprezentowano tekst bazowy (ang. *draft*) stworzony przez hipotetycznego badacza. Jak już wspomniano, zawiera on celowo wprowadzone deficyty w zakresie gramatyki, stylistyki oraz ortografii, co miało symulować naturalną strukturę roboczego szkicu przed ostateczną redakcją stworzonego przez człowieka:

> Tekst bazowy (ludzki): Bronisław „Borsuk" Wierzbicki był jednym z najbardziej enigmatycznych muzyków jazzowych w czasie odwilży i lat sześciesiątych. Był ceniony zarówno jako saksofonista tenorowy oraz jako kompozytor i zyskał status legendarnego tytana improwizacji, który łączył amerykański hard bop z subtelną słowiańską melancholią. Zapewniło mu to nieoficjalne miano „polskiego Coltrane'a". Miał zespół „Echa Północy" z którym koncertował na Festiwalu Jazz Jamboree i odnosił tam sukcesy. Oprócz tego koncertował w ograniczonym zakresie za żelazną kurtyną co stanowiło ewenement. Apogeum jego kariery zakończyło się zwrotem akcji, który był nagły i miał miejsce w 1968. Będąc w trasie koncertowej we Francji miały miejsce niejasne incydenty po których nie zdecydował się, jak wielu jego kolegów, na pozostanie na zachodzie, ale powrócił do kraju by w ciągu kolejnych kilku miesięcy kompletnie wycofać się z życia publicznego i porzucić muzykę oraz podjąć pracę w warszawie jako technik naprawy sprzętu RTV. Nie są znane motywy tej rezygnacji, a plotki w tym temacie dotyczą od szantażu przez SB po hipotezy o duchowym kryzysie, który miał być głęboki. Pozostają one do dzisiaj tematem do spekulacji historyków mózyki. Co kładzie cień na jego wcześniejszą spuściznę, która była genialna.

Przed realizacją zadań związanych z redakcją tekstu zweryfikowano bazę wiedzy wybranych do badania modeli językowych: ChatGPT oraz Claude Opus 4.1. Celem tej procedury było potwierdzenie, że modele nie dysponują wiedzą na temat fikcyjnej postaci Bronisława „Borsuka" Wierzbickiego, co stanowiło istotny element projektu badawczego. Uzyskane odpowiedzi od obydwóch modeli, potwierdzające brak jakiejkolwiek wiedzy na temat rzekomego jazzmana (zob. Rysunek 1 i Rysunek 2), zagwarantowały, że ich rola została ograniczona wyłącznie do umiejętności lingwistycznych i edycyjnych, a nie do generowania treści merytorycznej.


Rysunek. 1. Odpowiedź ChatGPT na pytanie o Bronisława „Borsuka" Wierzbickiego.

Źródło: opracowanie własne


Rysunek. 2. Odpowiedź Claude Opus 4.1
na pytanie o Bronisława „Borsuka" Wierzbickiego.

Źródło: opracowanie własne

Następnie tekst bazowy, zawierający celowo wprowadzone deficyty, posłużył jako dane wejściowe dla modeli ChatGPT oraz Claude Opus 4.1 z opcją użycia formalnego stylu wypowiedzi. W obu przypadkach zastosowano identyczne polecenie przedstawione poniżej.

> Treść polecenia wydanego modelom: Jako biegły korektor języka polskiego pracujący w wydawnictwie naukowym zredaguj poniższy tekst źródłowy, tak aby 1) nie utracić żadnych informacji w tekście wynikowym, 2) poprawić błędy ortograficzne, gramatyczne, interpunkcyjne i stylistyczne, 3) poprawić ogólny przepływ treści, tak aby tekst miał charakter formalny, a jednocześnie był łatwy w odbiorze przez potencjalnego czytelnika, 4) dla poprawy płynności wypowiedzi dostosuj szyk zdań i przenoś treści pomiędzy zdaniami, 5) nie używaj żadnego formatowania (pogrubienie, kursywa, listy numerowane/punktowane, *markdown*) – to ma być zwykły tekst, co najwyżej oddzielany akapitami.

Dla modelu ChatGPT po przetworzeniu tekstu wejściowego otrzymano następującą wersję zredagowaną:

> Tekst zredagowany przez ChatGPT: Bronisław „Borsuk" Wierzbicki był jednym z najbardziej enigmatycznych muzyków jazzowych okresu odwilży i lat sześćdziesiątych. Ceniono go zarówno jako saksofonistę tenorowego, jak i kompozytora. Zyskał legendarną opinię mistrza improwizacji, łączącego amerykański hard bop z subtelną, słowiańską melancholią. To właśnie dzięki temu zaczęto określać go mianem „polskiego Coltrane'a".
> Był liderem zespołu „Echa Północy", z którym występował na Festiwalu Jazz Jamboree, odnosząc tam znaczące sukcesy. Występował również, choć w ograniczonym zakresie, po obu stronach żelaznej kurtyny, co w tamtych czasach stanowiło ewenement.
> Apogeum jego kariery zakończyło się jednak nagłym i nieoczekiwanym zwrotem w roku 1968. Podczas trasy koncertowej we Francji doszło do niejasnych incydentów, po których Wierzbicki,

> w przeciwieństwie do wielu kolegów decydujących się na emigrację, powrócił do kraju. W ciągu kilku następnych miesięcy całkowicie wycofał się z życia publicznego, porzucił muzykę i podjął pracę w Warszawie jako technik naprawy sprzętu RTV.
> Motywy tej radykalnej decyzji pozostają nieznane. Krążyły różnorodne spekulacje – od domniemanego szantażu ze strony Służby Bezpieczeństwa po hipotezy o głębokim kryzysie duchowym. Do dziś pozostają one przedmiotem dyskusji historyków muzyki, rzucając cieniem na spuściznę artysty, choć ta bezsprzecznie pozostaje wybitna.

Dla modelu Claude Opus 4.1 proces redakcyjny zaowocował następującą wersją tekstu:

> Tekst zredagowany przez Claude Opus 4.1: Bronisław „Borsuk" Wierzbicki należał do grona najbardziej enigmatycznych muzyków jazzowych okresu odwilży i lat sześćdziesiątych. Jako saksofonista tenorowy oraz kompozytor zyskał uznanie i status legendarnego mistrza improwizacji, który potrafił łączyć amerykański hard bop z subtelną słowiańską melancholią. Dzięki temu unikatowemu stylowi otrzymał nieoficjalny tytuł „polskiego Coltrane'a".
> Na czele zespołu „Echa Północy" występował na prestiżowym Festiwalu Jazz Jamboree, odnosząc tam znaczące sukcesy. Ponadto koncertował w ograniczonym zakresie za żelazną kurtyną, co stanowiło w tamtych czasach prawdziwy ewenement. Apogeum jego kariery zakończyło się jednak nagłym zwrotem akcji w 1968 roku. Podczas trasy koncertowej we Francji doszło do niejasnych incydentów, po których Wierzbicki, w przeciwieństwie do wielu swoich kolegów, nie zdecydował się na pozostanie na Zachodzie i powrócił do kraju.
> W ciągu kilku kolejnych miesięcy artysta całkowicie wycofał się z życia publicznego, porzucił muzykę i podjął pracę w Warszawie jako technik naprawy sprzętu RTV. Motywy tej drastycznej rezygnacji pozostają nieznane, a krążące plotki obejmują hipotezy od szantażu ze strony Służby Bezpieczeństwa po teorie dotyczące głębokiego kryzysu duchowego. Spekulacje te do dziś stanowią przedmiot zainteresowania historyków muzyki, rzucając cień na wcześniejszą, genialną spuściznę artysty.

Ostatecznie autorski surowy tekst roboczy oraz obydwa teksty skorygowane przez DMJ zostały poddane ocenie przez usługę ZeroGPT.

## 4. Wyniki

Redakcja tekstu bazowego przeprowadzona przez model ChatGPT zaowocowała znaczną poprawą płynności, spójności oraz zgodności z normami języka formalnego. Zmiany miały charakter wyłącznie stylistyczno-gramatyczny, ściśle odpowiadający poleceniu wykluczającemu wprowadzanie nowych informacji merytorycznych.

Model dokonał korekty błędów formalnych i ortograficznych, eliminując literówki takie jak „sześciesiątych" (zmienione na „sześćdziesiątych"), „w warszawie" (na „w Warszawie") oraz „mózyki" (na „muzyki"), a także poprawiając nieprawidłową interpunkcję. Jedną z kluczowych interwencji była poprawa spójności i redukcja redundancji. Model wyeliminował tego typu konstrukcje, na przykład skracając zdanie dotyczące apogeum kariery: „Apogeum jego kariery zakończyło się zwrotem akcji, który był nagły i miał miejsce w 1968" do zwięzłego: „Apogeum jego kariery zakończyło się jednak nagłym i nieoczekiwanym zwrotem w roku 1968". DMJ wzmocnił spójność leksykalną przez zastąpienie powtórzeń i ogólników bardziej precyzyjnym słownictwem. Zmienił także nieformalne sformułowanie:

„Miał zespół «Echa Północy»" na bardziej formalne „Był liderem zespołu «Echa Północy»". Zmiany stylistyczne i syntaktyczne objęły podzielenie kilku długich, złożonych zdań na krótsze jednostki, zwiększając tym samym czytelność i przejrzystość tekstu. Wprowadzono inwersję składniową oraz zróżnicowano konstrukcje zdaniowe, a błędy gramatyczne i kolokacyjne zostały poprawione. Podsumowując, wersja zredagowana przez ChatGPT jest semantycznie tożsama z tekstem bazowym, ale wykazuje znacznie wyższą jakość lingwistyczną, charakteryzującą się zredukowaną perpleksją (większą przewidywalnością i płynnością) oraz wyższą spójnością syntaktyczną, odpowiadającą standardom profesjonalnej publikacji.

Model Claude Opus 4.1 wykazał równie wysoką zdolność do podniesienia jakości tekstu bazowego, co ChatGPT, lecz operował na nieco innych, choć komplementarnych, strategiach edycyjnych. Podobnie jak poprzedni model, Claude wyeliminował wszystkie błędy ortograficzne („sześciesiątych" na „sześćdziesiątych", „w warszawie" na „w Warszawie", „mózyki" na „muzyki") oraz zajął się korektą interpunkcyjną i gramatyczną. W zakresie syntaktyki, model znacząco poprawił strukturę otwierającą akapit. Na przykład, nadmiernie zagęszczone pierwsze zdanie tekstu bazowego zostało przekształcone w bardziej płynną sekwencję: „Bronisław «Borsuk» Wierzbicki należał do grona najbardziej enigmatycznych muzyków jazzowych okresu odwilży i lat sześćdziesiątych." oraz „Jako saksofonista tenorowy oraz kompozytor zyskał uznanie i status legendarnego mistrza improwizacji...".

W kontekście stylistyki, model Claude Opus 4.1 skupił się na wzmacnianiu formalności i precyzji leksykalnej. Zastąpił mniej formalne sformułowania, takie jak „Miał zespół «Echa Północy» z którym koncertował..." na „Na czele zespołu «Echa Północy» występował...". Model wyeliminował nadmierne przymiotniki, np. przekształcając „zwrotem akcji, który był nagły i miał miejsce w 1968" na zwięzłe „nagłym zwrotem akcji w 1968 roku". Zauważalne jest także ulepszenie klauzuli końcowej, gdzie „plotki w tym temacie dotyczą od szantażu przez SB po hipotezy o duchowym kryzysie, który miał być głęboki" zmieniono na bardziej formalne: „krążące plotki obejmują hipotezy od szantażu ze strony Służby Bezpieczeństwa po teorie dotyczące głębokiego kryzysu duchowego". Podsumowując, redakcja wykonana przez Claude Opus 4.1 nie zmienia tekstu pod względem merytorycznym, ale osiąga bardzo wysoki poziom płynności i spójności charakterystyczny dla profesjonalnej redakcji.

Chociaż obydwa modele osiągnęły tożsamy cel, jakim była eliminacja błędów formalnych i podniesienie płynności tekstu do standardu publikacyjnego, drobiazgowa analiza pozwala na zaobserwowanie subtelnych różnic w ich strategii redakcyjnej. Zarówno ChatGPT, jak i Claude Opus 4.1 z powodzeniem poprawiły błędy ortograficzne i gramatyczne oraz usunęły pleonazmy, redukując tym samym stopień perpleksji tekstu.

Model ChatGPT wydaje się w niektórych miejscach dążyć do większej zwięzłości i elegancji stylistycznej, wprowadzając np. frazę o „nagłym i nieoczekiwanym

zwrocie" zamiast oryginalnego, bardziej rozwlekłego opisu, lub zmieniając „legendarnego tytana improwizacji" na „legendarną opinię mistrza improwizacji". Natomiast Claude Opus 4.1 wykazał nieco większą tendencję do zachowania pierwotnej struktury semantycznej, dokonując delikatniejszych zmian, ale jednocześnie wprowadzając bardziej formalne i encyklopedyczne frazy, takie jak „należał do grona najbardziej enigmatycznych muzyków" oraz „otrzymał nieoficjalny tytuł".

Ostatecznie, o ile oba rezultaty są niemal perfekcyjne, tekst wygenerowany przez Claude Opus 4.1 sprawia wrażenie nieznacznie bardziej usystematyzowanego i wygładzonego pod kątem syntaktycznym i leksykalnym, osiągając ten „idealny" poziom płynności, który jest charakterystyczny dla maszynowego generowania. Ze względu na nieznaczną przewagę w zakresie spójności leksykalnej i formalnego tonu, można stwierdzić, że w tym konkretnym zadaniu redakcyjnym model Claude Opus 4.1 nieznacznie wyprzedza swojego konkurenta, dostarczając tekst o wyjątkowo niskiej perpleksji.

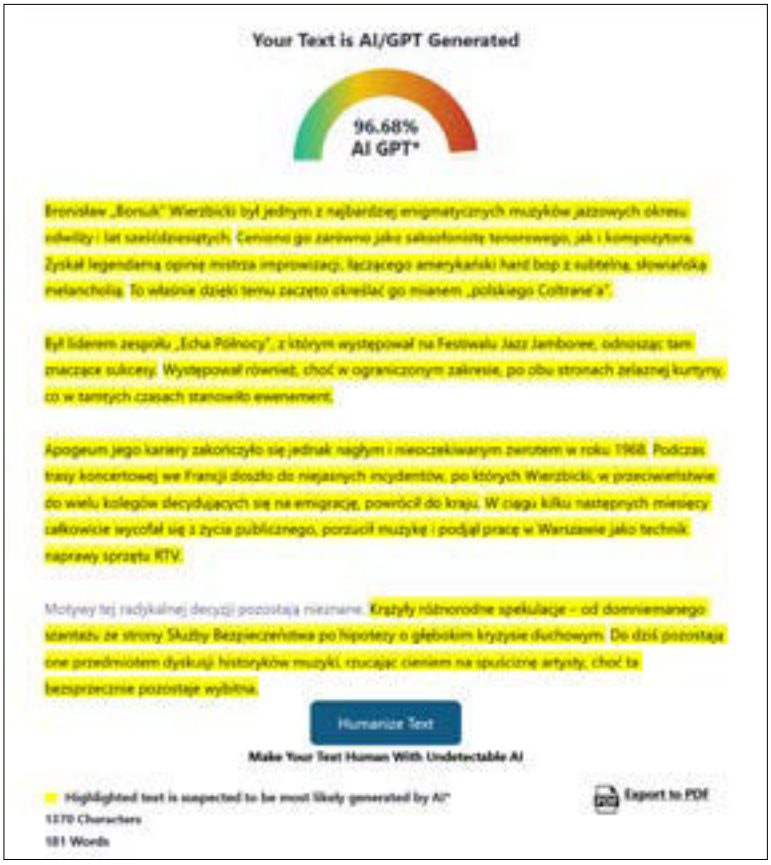

Rysunek 3. Wynik detekcji tekstu ludzkiego przez usługę internetową ZeroGPT.

Źródło: opracowanie własne

Subiektywna ocena, wskazująca na wyższą płynność i optymalizację tekstu przez model Claude Opus 4.1, znalazła swoje empiryczne potwierdzenie w pomiarach

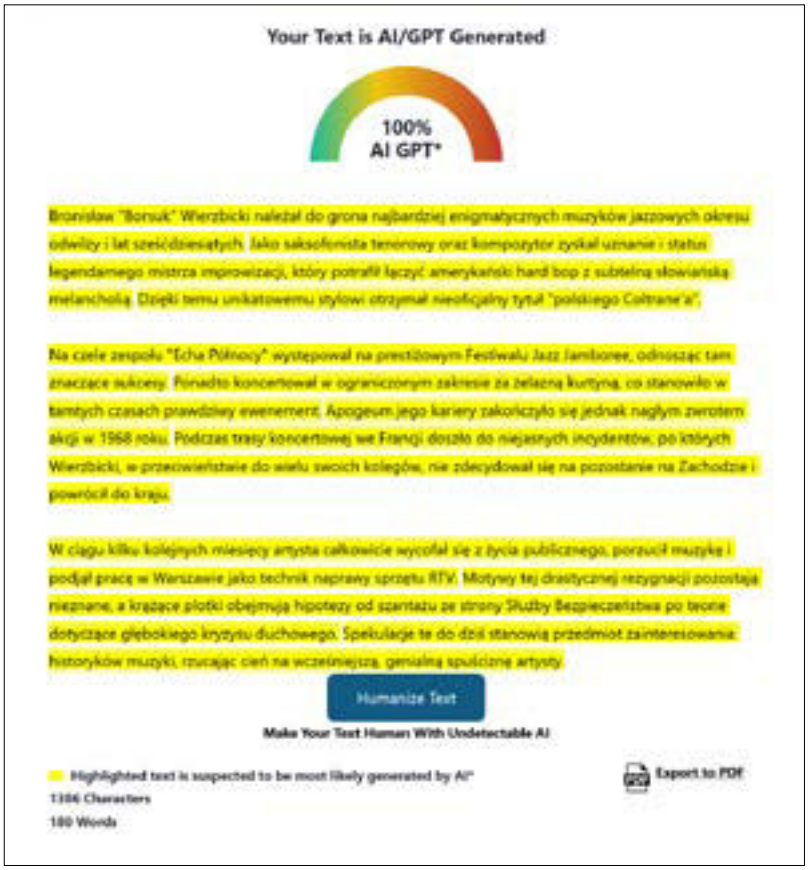
zrealizowanych za pomocą detektora treści SI ZeroGPT. Tekst bazowy, stanowiący roboczy szkic autorski, został prawidłowo zidentyfikowany przez narzędzie, osiągając 0% prawdopodobieństwa wygenerowania przez sztuczną inteligencję (zob. Rysunek 3). Tym samym zweryfikowano, że sygnatura lingwistyczna tekstu wejściowego była jednoznacznie ludzka.

Natomiast, mimo iż oba modele DMJ, zgodnie z instrukcją, pełniły wyłącznie rolę redakcyjną i nie wniosły żadnego wkładu merytorycznego, ich wpływ na syntaktykę i płynność tekstu natychmiast wywołał reakcję detektora. Wersja zredagowana przez model ChatGPT uzyskała wskaźnik wykrywalności na poziomie 96,7% (zob. Rysunek 4), podczas gdy tekst przetworzony przez model Claude Opus 4.1 został oznaczony z maksymalnym wynikiem 100% jako wygenerowany przez SI (zob. Rysunek 5).




Rysunek. 4. Wynik detekcji tekstu ChatGPT przez usługę internetową ZeroGPT.

Źródło: opracowanie własne




Rysunek 5. Wynik detekcji tekstu Claude przez usługę internetową ZeroGPT.

Źródło: opracowanie własne

Rezultaty te jednoznacznie dowodzą, że detektor nie identyfikuje pochodzenia intelektualnego treści (wkładu merytorycznego), lecz koncentruje się wyłącznie na stylistycznych i syntaktycznych cechach tekstu (niskiej perpleksji), będących efektem optymalizacji językowej.

## 5. Dyskusja

Wyniki przeprowadzonego studium przypadku dostarczają jednoznacznej odpowiedzi na pierwsze pytanie badawcze: tekst, którego autorem merytorycznym w 100% jest człowiek, może zostać błędnie zakwalifikowany jako dzieło SI wyłącznie po przeprowadzeniu na nim korekty językowej za pomocą DMJ. Fundamentalną konkluzją jest zatem fakt, że obecne narzędzia do detekcji treści generowanych

przez sztuczną inteligencję nie są w stanie odróżnić autorstwa intelektualnego od redakcji stylistycznej. Algorytmy te, zamiast analizować pochodzenie merytorycznej treści, skupiają się na identyfikacji wzorców językowych korelujących z wysoką jakością redakcyjną – płynnością, spójnością i niską perpleksją. Paradoksalnie, cechy te są nie tylko sygnaturą tekstów generowanych maszynowo, ale również celem, do którego dąży każdy autor pragnący stworzyć klarowny i profesjonalny tekst naukowy. W konsekwencji, im skuteczniej autor – wspomagany przez zaawansowane narzędzie – poprawi swój tekst, tym większe prawdopodobieństwo, że zostanie on błędnie oznaczony jako dzieło maszyny.

Prowadzi to wprost do drugiego pytania badawczego, dotyczącego granicy między etycznym wykorzystaniem SI a nieuczciwością akademicką. Na podstawie przeprowadzonych analiz należy stwierdzić, że granica ta nie leży w samym fakcie użycia narzędzia, lecz w intencji i sposobie jego zastosowania. Wykorzystanie DMJ jako zaawansowanego korektora, którego rola jest analogiczna do ludzkiego redaktora, jest działaniem w pełni etycznym. Nieuczciwość akademicka rozpoczyna się w momencie, gdy autor ceduje na maszynę odpowiedzialność za wkład intelektualny – generowanie tez, argumentów, analizę danych czy wnioskowanie – a następnie przypisuje sobie ich autorstwo. Kluczem jest zatem pochodzenie idei, a nie narzędzie użyte do ich językowego udoskonalenia.

Bezkrytyczne poleganie na detektorach SI, które tej granicy nie rozpoznają, rodzi szereg negatywnych konsekwencji dla całego ekosystemu naukowego, co stanowi odpowiedź na trzecie pytanie badawcze. Dla studentów i badaczy oznacza to absurdalny wybór między dążeniem do doskonałości językowej a ryzykiem fałszywych oskarżeń. Dla nauczycieli, recenzentów i redaktorów – pokusa użycia detektora jako szybkiego i pozornie obiektywnego arbitra może prowadzić do niesprawiedliwych ocen i odrzucania wartościowych prac. W szerszej perspektywie, instytucje naukowe, tworząc polityki oparte na zawodnych wskaźnikach, zamiast budować kulturę zaufania, promują atmosferę podejrzliwości, co prowadzi do patologicznej „gry w kotka i myszkę" z narzędziami do „humanizacji" tekstów.

Na ten obraz nakłada się dodatkowa warstwa problematyczna, związana z etycznym i prawnym statusem samych usług detekcyjnych. Wiele z tych platform działa na podstawie niejasnych warunków użytkowania, nie dając gwarancji, co dzieje się z przesyłanymi do nich treściami. Przekazywanie nieopublikowanych prac naukowych, esejów studenckich czy danych badawczych do zewnętrznych podmiotów komercyjnych rodzi poważne ryzyko naruszenia praw autorskich, tajemnicy badawczej czy po prostu utraty kontroli nad własnością intelektualną. Sytuacja staje się jeszcze bardziej skomplikowana, gdy osoba weryfikująca tekst (np. wykładowca czy recenzent) nie jest jego autorem, co może prowadzić do naruszenia praw osób trzecich.

## 6. Wnioski

Przeprowadzone badanie, mimo swojego pilotażowego charakteru, dostarcza empirycznych dowodów na to, że wysoki wskaźnik prawdopodobieństwa wygenerowania tekstu przez SI nie może być traktowany jako wiarygodny dowód nieuczciwości akademickiej. Wykorzystanie dużego modelu językowego w roli cyfrowego redaktora jest uprawnionym i etycznie dopuszczalnym wsparciem w procesie twórczym, a ostateczna odpowiedzialność za treść, jej prawdziwość i oryginalność myśli zawsze spoczywa na człowieku. Głównym przesłaniem płynącym z niniejszej pracy jest zatem konieczność fundamentalnej zmiany paradygmatu: od pytania „Czy napisała to sztuczna inteligencja?" do pytania „Czy autor w pełni ręczy za tę pracę?" Promowanie transparentności w zakresie używanych narzędzi, w połączeniu z edukacją na temat ich możliwości i ograniczeń, jest znacznie bardziej konstruktywnym podejściem niż budowanie polityk opartych na zawodnych i łatwych do zmanipulowania technologiach detekcyjnych.

Należy przy tym podkreślić, że choć artykuł koncentruje się wyłącznie na wąskim, redakcyjnym zastosowaniu GenSI, problem ten należy postrzegać w znacznie szerszym kontekście. Podsycanie w środowiskach akademickich zjawiska „GENgrozy" – lęku przed adopcją technologii generatywnej sztucznej inteligencji – stwarza realne ryzyko dla konkurencyjności polskiej nauki na arenie międzynarodowej.

Podczas gdy polskie instytucje wciąż debatują nad elementarnymi zasadami, świat nauki nie czeka. Dopuszczanie GenSI do szerokiego spektrum zadań badawczych staje się normą. Czołowe globalne czasopisma naukowe odchodzą od prohibicyjnych regulacji, koncentrując się na transparentności (Wang and Zhao, 2024): wymagają od autorów jedynie klarownej deklaracji zakresu użycia SI, analogicznie do sposobu, w jaki deklaruje się wkład poszczególnych autorów w pracach wieloautorskich. Co więcej, wiodące uczelnie zagraniczne aktywnie tworzą kompleksowe polityki etyki użycia AI, precyzyjnie wyznaczając granice pomiędzy dozwolonym, wspierającym zastosowaniem technologii w badaniach i dydaktyce a przekroczeniem granic etycznych. Nawet kluczowa infrastruktura akademicka integruje te zmiany: dostawcy naukowych baz danych, takich jak Scopus czy Web of Science, wdrażają technologie GenSI bezpośrednio w interfejsy wyszukiwawcze.

Tymczasem to opóźnienie staje się problemem także na własnym podwórku: nawet na poziomie krajowym, zarówno w sektorach komercyjnych, jak i w administracji publicznej, obserwujemy znaczący postęp. Świadczy o tym opracowanie i ciągłe doskonalenie krajowych dużych modeli językowych (jak Bielik i PLLuM) oraz opracowywanie polityk i pilotażowe integracje tych technologii z infrastrukturą rządową – nie tylko centralną, ale i samorządową. Wykluczenie i piętnowanie GenSI przez polską naukę w tym momencie może nieść za sobą znacznie głębsze konsekwencje. Rola nauki w wolnorynkowych i demokratycznych ustrojach polega na służeniu społeczeństwu, a nie służeniu samej sobie. Jeżeli środowiska naukowe

będą udawać, że omawiany tutaj zwrot technologiczny ich nie dotyczy, postawią pod znakiem zapytania społeczne uzasadnienie swojego istnienia i swoją zdolność do bycia katalizatorem krajowego postępu.

Wracając jednak do *stricte* redakcyjnego zastosowania GenSI, należy wskazać potencjalnie najgorsze efekty tej moralnej paniki. Prawdziwymi beneficjentami obecnej sytuacji stają się bowiem jedynie dostawcy internetowych usług „humanizujących" teksty, które pozwalają na oszukanie detektorów SI. Należy mieć świadomość, że długofalowe skutki masowego korzystania z takich platform są potencjalnie wysoce negatywne. Teksty celowo „psute" w celu symulowania ludzkich niedoskonałości będą stopniowo zasilać przyszłe korpusy danych treningowych, co w perspektywie czasu może prowadzić do degradacji dużych modeli językowych i erozji jakości języka naturalnego w przestrzeni cyfrowej. Wydaje się, że skutecznym antidotum na wspomnianą „GENgrozę" jest opracowywanie polityk etycznego, ale i aktywnego integrowania technologii GenSI w procesy dydaktyczne i badawcze przez poszczególne jednostki akademickie.

## Bibliografia:

Barrot, J. S. (2020). Integrating technology into ESL/EFL writing through Grammarly. *RELC Journal*, *53*(3), 764–768. doi: 10.1177/0033688220966632

Francis, N. J., Jones, S., Smith, D. P. (2025). Generative AI in higher education: Balancing innovation and integrity. *British Journal of Biomedical Science*, *81*(14048), 1–9. doi: 10.3389/bjbs.2024.14048

Helgesson, G., Eriksson, S. (2015). Plagiarism in research. *Medicine, Health Care and Philosophy*, *18*, 91–101. doi: 10.1007/s11019–014–9583–8

Luo, J. (2024). A critical review of GenAI policies in higher education assessment: A call to reconsider the "originality" of students' work. *Assessment & Evaluation in Higher Education*, *49*(5), 651–664. doi: 10.1080/02602938.2024.2309963

Oakman, R. L. (1994). The evolution of intelligent writing assistants: trends and future prospects. In: *Proceedings Sixth International Conference on Tools with Artificial Intelligence. TAI 94* (233–234). Institute of Electrical and Electronics Engineer. doi: 10.1109/TAI.1994.346488

Tillmanns, T., Filho, A. S., Rudra, S., Weber, P., Dawitz, J., Wiersma, E., Dudenaite, D., Reynolds, S. (2025). Mapping tomorrow's teaching and learning spaces: A systematic review on GenAI in higher education. *Trends in Higher Education*, *4*(1), 2. doi: 10.3390/higheredu4010002

Wang, Y., Zhao, L. (2024). Toward the transparent use of Generative Artificial Intelligence in academic articles. *Journal of Scholarly Publishing, 55*(4), 467–484. doi: 10.3138/jsp-2023–0053

Yusuf, A., Pervin, N., Román-González, M., Noor, N. M. (2024). Generative AI in education and research: A systematic mapping review. *Review of Education*, *12*(2), e3489. doi: 10.1002/rev3.3489

# Has AI written this paper? A case study in „plagiarism detection"

**Abstract**

**Purpose/Thesis:** The paper evaluates the use of GenAI content detectors in the context of verifying academic integrity. The main thesis is that content detectors cannot distinguish between AI-generated text and human-authored content that has only been stylistically edited by large language models. Consequently, high detection scores do not constitute reliable evidence of plagiarism.

**Approach/Methods:** A case study method was employed. A biography of a fictional character was created by a human and subsequently edited using ChatGPT and Claude with precise instructions limiting the AI to a purely editorial role. All versions were analysed using the ZeroGPT detector.

**Results and conclusion:** The original text scored a 0% probability of being AI-generated, while the versions edited by ChatGPT and Claude achieved 96.7% and 100%, respectively. Detectors identify stylistic features characteristic of well-edited texts, rather than the actual origin of the intellectual contribution.

**Practical implications:** The findings indicate the need to revise academic policies in light of AI detectors. Institutions should promote transparent rules for using AI as a supporting tool instead of relying on unreliable technologies that lead to unfair assessments.

**Originality/Value:** The work provides empirical evidence of the unreliability of AI detectors in the context of plagiarism detection, establishes a clear ethical boundary between using AI as a tool and academic dishonesty, and promotes a conscious approach to new technologies in academia.

**Keywords**

AI content detection. Case study. Editorial tools. Ethics in science. Generative artificial intelligence (GenAI). Large language models (LLM). Plagiarism. Scholarly communication.

*ANNA MAŁGORZATA KAMIŃSKA jest adiunktem w Instytucie Nauk o Kulturze Uniwersytetu Śląskiego w Katowicach oraz pracownikiem Biblioteki Politechniki Śląskiej w Gliwicach. Jej zainteresowania naukowe obejmują informatologię, naukometrię, bibliologię, komunikację naukową oraz zagadnienia związane z przetwarzaniem języka naturalnego. W badaniach koncentruje się na analizie cytowań, bazach bibliograficznych i zrównoważonym rozwoju w bibliotekoznawstwie.*

*Kontakt:*
*anna.kaminska@us.edu.pl*
*Uniwersytet Śląski w Katowicach*
*Wydział Humanistyczny*
*Instytut Nauk o Kulturze*
*ul. Uniwersytecka 4, 40-007 Katowice*