ORIGINALNE BADANIA

Duże modele językowe: pionierskie nowe horyzonty edukacyjne w dziedzinie krótkowzroczności dziecięcej

Mohammad Delsoz · Amr Hassan · Amin Nabavi · Amir Rahdar · Brian Fowler · Natalie C. Kerr · Lauren Claire Ditta · Mary E. Hoehn · Margaret M. DeAngelis · Andrzej Grzybowski · Yih-Chung Tham · Siamak Yousefi

Otrzymano: 12 lutego 2025 r. / Przyjęto: 2 kwietnia 2025 r.

STRESZCZENIE

***Wprowadzenie*:** Celem niniejszego badania była ocena wydajności trzech dużych modeli językowych (LLM), a mianowicie ChatGPT-3.5, ChatGPT-4o (o1 Preview) i Google Gemini, w zakresie tworzenia materiałów edukacyjnych dla pacjentów (PEM) oraz Poprawa czytelności internetowych PEM dotyczących krótkowzroczności dziecięcej na stronie .

***Metody*:** Odpowiedzi wygenerowane przez LLM zostały ocenione przy użyciu trzech poleceń. Polecenie A brzmiało: „Napisz materiał edukacyjny na temat krótkowzroczności u dzieci". Polecenie B zawierało dodatkowy modyfikator określający „poziom czytania dla szóstej klasy według formuły czytelności FKGL (Flesch-Kincaid Grade Level)". Wskazówka C miała na celu przepisanie istniejących materiałów edukacyjnych dla pacjentów (PEM) na poziom szóstej klasy przy użyciu FKGL. Odpowiedzi zostały ocenione pod kątem jakości (narzędzie DISCERN), czytelności (FKGL, SMOG (Simple Measure of Gobbledygook)), materiałów edukacyjnych dla pacjentów

Informacje dodatkowe Wersja internetowa zawiera materiały dodatkowe dostępne pod adresem https://doi.org/10.1007/s40123-025-01142-x.

M. Delsoz · A. Nabavi · A. Rahdar · B. Fowler · N. C. Kerr · L. C. Ditta · M. E. Hoehn · S. Yousefi (✉)
Hamilton Eye Institute, Wydział Okulistyki, Centrum Nauk o Zdrowiu Uniwersytetu Tennessee, 930 Madison Ave., Suite 471, Memphis, TN 38163, USA
e-mail: Siamak.Yousefi@uthsc.edu

A. Hassan
Katedra Okulistyki, Instytut Okulistyki im. Gavina Herberta, Uniwersytet Kalifornijski, Irvine, Kalifornia, USA

M. M. DeAngelis
Katedra Okulistyki, Uniwersytet w Buffalo, Buffalo, NY, USA

M. M. DeAngelis
Służba badawcza, VA Western New York Healthcare System, Buffalo, NY, USA

A. Grzybowski
Instytut Badań Okulistycznych, Fundacja Rozwoju Okulistyki, Poznań, Polska

Y.-C. Tham
Szkoła Medyczna Yong Loo Lin, Narodowy Uniwersytet Singapuru, Singapur, Republika Singapuru

Y.-C. Tham
Centrum Innowacji i Precyzyjnej Opieki Okulistycznej, Wydział Okulistyki, Yong Loo Lin School of Medicine, Narodowy Uniwersytet Singapuru i Narodowy System Opieki Zdrowotnej, Singapur, Republika Singapuru

S. Yousefi
Wydział Genetyki, Genomiki i Informatyki, Centrum Nauk o Zdrowiu Uniwersytetu Tennessee, Memphis, TN, USA

Published online: 21 April 2025

Narzędzie oceny (PEMAT, zrozumiałość/ wykonalność) i dokładność.
***Wyniki*:** ChatGPT-4o (01) i ChatGPT-3.5 wygenerowały dobrej jakości PEM (odpowiednio DISCERN 52,8 i 52,7); jednak jakość spadła od podpowiedzi A do podpowiedzi B (p= 0,001 i p= 0,013). Google Gemini wygenerowało teksty o średniej jakości (DISCERN 43), ale poprawiło się w przypadku promptu B (p= 0,02). Wszystkie PEM przekroczyły próg zrozumiałości PEMAT wynoszący 70%, ale nie osiągnęły progu wykonalności wynoszącego 70% (40%). Nie zidentyfikowano żadnych błędnych informacji. Czytelność poprawiła się po zastosowaniu podpowiedzi B; ChatGPT-4o (01) i ChatGPT-3.5 osiągnęły poziom szóstej klasy lub niższy (FGKL 6± 0,6 i 6,2± 0,3), podczas gdy Google Gemini osiągnął
nie (FGKL 7± 0,6). ChatGPT-4o (01) przewyższył Google Gemini pod względem czytelności (p< 0,001), ale był porównywalny z ChatGPT-3.5 (p= 0,846). Prompt C poprawił czytelność we wszystkich modelach LLM, przy czym ChatGPT-4o (o1 Preview) wykazał najbardziej znaczący wzrost (FKGL 5,8± 1,5; p< 0,001).
***Wnioski*:** ChatGPT-4o (o1 Preview) wykazuje potencjał w tworzeniu dokładnych, dobrej jakości i zrozumiałych materiałów edukacyjnych dla pacjentów oraz w ulepszaniu internetowych materiałów edukacyjnych dotyczących krótkowzroczności u dzieci.

Słowa kluczowe: duże modele językowe; materiały edukacyjne dla pacjentów; krótkowzroczność u dzieci

Najważniejsze punkty podsumowania

W niniejszym badaniu oceniono zdolność narzędzi opartych na sztucznej inteligencji (AI), w tym ChatGPT-3.5, ChatGPT-4o (o1 Preview) i Google Gemini, do generowania i poprawiania czytelności istniejących materiałów edukacyjnych dotyczących krótkowzroczności u dzieci.

Wyniki badania pokazują, że modele te, w szczególności ChatGPT-4o (o1 Preview), mogą generować treści dobrej jakości, zrozumiałe, dokładne, przyjazne dla użytkownika, spełniające standardy czytelności oraz poprawiające czytelność istniejących materiałów edukacyjnych online dotyczących krótkowzroczności u dzieci na poziomie szóstej klasy lub niższym.

Badanie podkreśla również znaczenie dostosowywania podpowiedzi w celu zwiększenia przejrzystości treści i zawiera praktyczne zalecenia dotyczące wykorzystania narzędzi AI do poprawy edukacji zdrowotnej.

W badaniu wskazano obszary wymagające poprawy, takie jak zwiększenie praktycznej użyteczności i włączenie elementów multimedialnych.

WPROWADZENIE

Krótkowzroczność dziecięca, powszechnie nazywana krótkowzrocznością, stała się poważnym problemem zdrowia publicznego na całym świecie. Najnowsze badania wskazują, że około jedna trzecia dzieci i młodzieży na całym świecie cierpi na krótkowzroczność, a prognozy szacują, że do 2050 r. częstość występowania tej wady wzrośnie do prawie 40%, co odpowiada ponad 740 milionom przypadków w tej grupie wiekowej [1]. Częstość występowania wśród dzieci jest szczególnie niepokojąca; badania wykazały, że w krajach Azji Wschodniej nawet 80–90% młodych dorosłych jest krótkowzrocznych [2]. W Stanach Zjednoczonych częstość występowania krótkowzroczności u dzieci znacznie wzrosła w ciągu ostatnich dziesięcioleci [3].

Wzrost częstości występowania krótkowzroczności u dzieci wiąże się z różnymi czynnikami, w tym wpływami środowiskowymi i predyspozycjami genetycznymi [4]. Warto zauważyć, że pandemia COVID-19 i związane z nią ograniczenia pogłębiły ten problem, zwiększając czas spędzany przed ekranem i ograniczając aktywność na świeżym powietrzu, które są istotnymi czynnikami ryzyka rozwoju i postępu krótkowzroczności [1]. Ta zmiana zachowań doprowadziła do znacznego wzrostu liczby przypadków krótkowzroczności wśród dzieci w trakcie pandemii i po jej zakończeniu. Skuteczne materiały edukacyjne dla pacjentów (PEM) mają kluczowe znaczenie w leczeniu i łagodzeniu tego schorzenia, ponieważ mogą one zwiększyć społeczne zrozumienie krótkowzroczności i zachęcić do proaktywnych zachowań zdrowotnych. Skuteczne leczenie i łagodzenie krótkowzroczności u dzieci w dużym stopniu opiera się na kompleksowych materiałach edukacyjnych dla pacjentów [5]. Zasoby te mają kluczowe znaczenie dla zwiększenia zrozumienia i promowania proaktywnych zachowań zdrowotnych wśród pacjentów i ich

rodzin. Jednak istniejące PEM często przekraczają poziom czytania zalecany przez Amerykańskie Towarzystwo Medyczne (AMA) dla uczniów szóstej klasy, co sprawia, że są one mniej dostępne dla ogółu społeczeństwa [6–8].

W ostatnim czasie pojawienie się dużych modeli językowych (LLM), w tym ChatGPT firmy OpenAI, Gemini firmy Google i innych chatbotów opartych na sztucznej inteligencji (AI), wprowadziło innowacyjne zastosowania w medycynie, zwłaszcza w okulistyce. Ostatnie badania wykazały obiecujące wyniki, pokazując, że ich biegłość w przetwarzaniu języka naturalnego umożliwia im interpretację złożonych danych medycznych, pomoc w diagnozowaniu schorzeń oczu, pomoc w badaniach naukowych oraz dostarczanie spersonalizowanych zaleceń dotyczących leczenia [9–15]. Modele oparte na sztucznej inteligencji wykazały potencjał w sektorze informacji zdrowotnej w różnych dziedzinach medycyny. Jednak ich skuteczność w tworzeniu PEM dostosowanych specjalnie do krótkowzroczności u dzieci oraz w poprawianiu czytelności istniejących PEM pozostaje niedostatecznie zbadana.

Celem niniejszego badania jest ocena jakości, czytelności, przydatności i dokładności modeli PEM dotyczących krótkowzroczności u dzieci, wygenerowanych przez modele LLM, w tym ChatGPT-3.5, ChatGPT-4o (wersja 01 Preview) i Google Gemini. Oceniając zdolność tych modeli do generowania zrozumiałych i przydatnych informacji zdrowotnych oraz poprawy czytelności istniejących zasobów internetowych, staramy się określić ich przydatność jako narzędzi uzupełniających w edukacji pacjentów.

METODY

Badanie zostało zwolnione z oceny etycznej Centrum Nauk o Zdrowiu Uniwersytetu Tennessee, ponieważ nie obejmowało uczestników ani ich danych osobowych, a skupiało się na ocenie działania najnowszych modeli sztucznej inteligencji. Skupienie się na publicznie dostępnych danych i tekstach generowanych przez sztuczną inteligencję zapewniło zgodność z normami dotyczącymi prywatności i etyki badań naukowych. Badanie zostało przeprowadzone w okresie od października do grudnia 2024 r., zgodnie z zasadami Deklaracji Helsińskiej.

Projekt badania

Celem badania była ocena przydatności modeli LLM w tworzeniu materiałów edukacyjnych dotyczących krótkowzroczności u dzieci oraz poprawie czytelności obecnych zasobów internetowych. Porównano wydajność modeli Chat-GPT-3.5, ChatGPT-4o (o1 Preview) i Gemini w generowaniu nowych materiałów edukacyjnych oraz przepisywaniu istniejących, aby uczynić je bardziej przystępnymi (dodatek 1 w materiałach uzupełniających).

Wybór dużych modeli językowych

W niniejszym badaniu oceniono modele LLM: Chat-GPT-3.5, ChatGPT-4o (o1 Preview) oraz Gemini. Modele te wybrano ze względu na ich powszechną dostępność oraz sprawdzoną zdolność do generowania i udoskonalania tekstu. Dostęp do modeli ChatGPT uzyskano za pośrednictwem platformy OpenAI, a do modelu Gemini – poprzez dedykowany interfejs internetowy.

Projektowanie promptów

W celu wygenerowania PEM dotyczących krótkowzroczności u dzieci opracowano dwa różne polecenia:

- Wskazówka A (kontrola): Napisz materiał edukacyjny na temat krótkowzroczności u dzieci, który będzie łatwo zrozumiały dla przeciętnego Amerykanina. Ta ogólna wskazówka została celowo opracowana bez konkretnych wytycznych dotyczących czytelności, aby ustalić podstawowe zrozumienie nieodłącznej czytelności i jakości PEM generowanych przez LLM.
- Wskazówka B (zmodyfikowana): „Ponieważ przeciętny Amerykanin ma poziom czytania odpowiadający szóstej klasie, wykorzystując formułę czytelności FKGL (Flesch-Kincaid Grade Level), napisz materiał edukacyjny na temat krótkowzroczności u dzieci, który będzie łatwy do zrozumienia dla przeciętnego Amerykanina” (zgodnie z zaleceniem Amerykańskiego Stowarzyszenia Medycznego) (Dodatek 2 w materiałach uzupełniających). To zadanie wyraźnie instruowało modele, aby generowały materiały odpowiednie dla poziomu czytania szóstoklasisty, odwołując się bezpośrednio do formuły czytelności FKGL. Wybraliśmy to szczegółowe sformułowanie, aby ocenić, jak skutecznie modele LLM mogą tworzyć ukierunkowane,

czytelne treści, gdy otrzymają wyraźne wytyczne, zgodnie z zaleceniami standardów AMA dotyczących czytelności PEM.

W punkcie B odwołaliśmy się do formuły czytelności Flesch-Kincaid Grade podczas tworzenia materiałów edukacyjnych dla pacjentów na poziomie dostosowanym do każdego LLM, głównie dlatego, że jest ona powszechnie uznawana i popierana przez znane organizacje, w tym AMA i National Institutes of Health (NIH) [16, 17], do oceny PEM. W przeciwieństwie do innych wskaźników czytelności, FKGL ocenia w szczególności długość zdań i złożoność słów, dwa kluczowe elementy wpływające na zrozumienie, dzięki czemu nadaje się szczególnie do informacji zdrowotnych skierowanych do szerokiego grona odbiorców. Jest to cenne narzędzie, zwłaszcza w dziedzinie edukacji i publikacji, zapewniające jasność i dostępność informacji dla docelowych odbiorców [17].

Każdy LLM miał za zadanie odpowiedzieć na te pytania (A i B) w 20 próbach, zapewniając oddzielne i bezstronne instancje czatu dla każdej próby, aby uniknąć wpływu funkcji uczenia się przez wzmocnienie na podstawie informacji zwrotnych od ludzi (RLHF) w chatbotach LLM [18, 19].

W celu ulepszenia już dostępnych zasobów internetowych, pierwsze 20 kwalifikujących się istniejących internetowych materiałów edukacyjnych dotyczących krótkowzroczności u dzieci zostało pozyskanych z dwóch pierwszych stron wyszukiwarki Google przy użyciu słowa kluczowego „krótkowzroczność u dzieci” w okresie od października do grudnia 2024 r. Zdajemy sobie sprawę, że ograniczenie wyszukiwania do pierwszych 20 kwalifikujących się wyników wyszukiwania Google może wprowadzać błąd selekcji, ponieważ wyniki te mogą nie reprezentować wszystkich istniejących zasobów internetowych. Jednak nasze uzasadnienie dla zastosowania tego podejścia wynika z rzeczywistych zachowań użytkowników. Liczne badania wskazują, że większość osób poszukujących informacji zdrowotnych w Internecie rzadko wychodzi poza dwie pierwsze strony wyników wyszukiwania, ponieważ stanowią one ponad 95% całkowitego ruchu internetowego wśród użytkowników wyszukiwarek [20, 21]. Skupiając się na materiałach, z którymi pacjenci mają największe szanse się spotkać, chcieliśmy ocenić treści edukacyjne, które mają największy praktyczny wpływ na wiedzę zdrowotną i edukację pacjentów. Chociaż zdajemy sobie sprawę, że nie obejmuje to wszystkich możliwych zasobów, odzwierciedla to źródła, z którymi pacjenci mają statystycznie większe szanse się spotkać.

do obejrzenia. Wykluczone materiały obejmowały reklamy, artykuły naukowe, rozdziały z książek, treści multimedialne, takie jak filmy, osobiste blogi, strony internetowe w językach innych niż angielski, zasoby dotyczące podejmowania decyzji klinicznych oraz źródła nieprzeznaczone dla pacjentów. Każdy PEM został wprowadzony do modeli LLM wraz z następującym komunikatem:

Wskazówka C: „Ponieważ materiały edukacyjne dla pacjentów (PEM) powinny być napisane na poziomie czytania szóstej klasy, czy mógłbyś przepisać poniższy tekst, aby spełniał ten standard, używając formuły czytelności FKGL? [wstaw tekst]”. Pytanie C ma na celu ocenę zdolności modeli LLM do poprawiania istniejących materiałów; w tym pytaniu modele otrzymały polecenie przepisania istniejących materiałów edukacyjnych dla pacjentów (PEM) tak, aby spełniały one wymagania poziomu czytania dla szóstej klasy, przy użyciu FKGL. Wyraźne odniesienie do FKGL ponownie pozwoliło nam ocenić biegłość modeli w dostosowywaniu złożoności istniejących treści do poziomu docelowego pacjentów, odzwierciedlając rzeczywiste scenariusze, w których istniejące materiały edukacyjne dla pacjentów (PEM) muszą być dostosowane w celu poprawy czytelności. Wreszcie, nasze podejście polegało na zebraniu 20 materiałów edukacyjnych PEM na każdy model LLM (łącznie 60) w poleceniu A, wygenerowaniu kolejnych 20 na każdy model LLM w poleceniu B (kolejne 60) oraz ocenie 20 istniejących zasobów edukacyjnych online i ulepszeniu 20 istniejących zasobów edukacyjnych online na trzy modele LLM (łącznie 60). W rezultacie przeanalizowano łącznie 200 materiałów edukacyjnych. Wykorzystaliśmy bezpłatne narzędzie internetowe readabilityformulas.com do obliczania wyników czytelności i innych kluczowych wskaźników [22]. Oprogramowanie to wykorzystuje sprawdzone narzędzia do obliczania czytelności i jest szeroko stosowane w renomowanych badaniach medycznych [23–25]. Wszystkie materiały edukacyjne PEM zostały poddane ocenie pod kątem podstawowych wskaźników czytelności, takich jak liczba sylab, liczba słów, liczba słów złożonych, liczba zdań, wskaźnik czytelności SMOG (Simple Measure of Gobbledygook) oraz wskaźnik FKGL.

Metryki oceny wygenerowanych PEM

Do oceny jakości i wiarygodności wygenerowanych i przeredagowanych PEM wykorzystaliśmy pełny, 16-punktowy, zatwierdzony instrument DISCERN [26]. To narzędzie oceny zawiera 16-punktowy kwestionariusz przeznaczony do użytku przez

recenzentów. Każda pozycja jest oceniana w skali od 1 do 5, gdzie wyniki 1–2 oznaczają niską jakość, 3 oznacza średnią jakość, a 4–5 wskazuje na wysoką jakość [27]. Kwestionariusz DISCERN składa się z trzech części. Pierwsza część (pytania 1–8) koncentruje się na wiarygodności publikacji. Ocenia takie aspekty, jak cel publikacji, wiarygodność źródeł informacji, jej trafność oraz to, czy zawiera dodatkowe zasoby dotyczące danego schorzenia. Druga sekcja (pytania 9–15) dotyczy szczegółowych informacji na temat metod leczenia, w tym ich działania, potencjalnych zagrożeń i korzyści oraz tego, czy przedstawiono alternatywne opcje leczenia. Trzecia sekcja (pytanie 16) ocenia ogólną jakość publikacji jako źródła informacji na temat leczenia (dodatek 3A w materiałach uzupełniających). Obliczyliśmy całkowity wynik DISCERN, sumując wszystkie 16 ocen (w zakresie 16–80) dla każdego recenzenta dla każdego materiału edukacyjnego. Następnie przyporządkowaliśmy wynik DISCERN do odpowiedniego poziomu jakości (tabela 1), zgodnie z zalecanymi metodami obliczeniowymi, i przeanalizowaliśmy go. System ocen, przedstawiony w tabeli 1, został opracowany na podstawie wyników uzyskanych przez poszczególne PEM [28]. Dwóch oceniających (okulistów) oceniło wygenerowane i przepisane materiały edukacyjne dla pacjentów (PEM) w postaci 120 ulotek. Aby zminimalizować stronniczość, oceniający nie znali ocen innych oceniających, usunięto tożsamość źródeł generujących odpowiedzi, a ostateczny wynik dla każdego materiału edukacyjnego dla pacjentów (PEM) ustalono na podstawie mediany wyników recenzentów. Ponadto oceniający ocenili zrozumiałość i wykonalność materiałów edukacyjnych dla pacjentów (PEM) za pomocą narzędzia Patient Education Materials Assessment Tool (PEMAT), zatwierdzonego instrumentu Agencji ds. Badań i Jakości Opieki Zdrowotnej (AHRQ) [29, 30] (dodatek 3B w materiałach uzupełniających). Zrozumiałość, definiowana jako zdolność różnych odbiorców do zrozumienia i wyjaśnienia głównych przekazów, została obliczona jako ogólny odsetek na podstawie 12 pytań typu tak/nie, obejmujących takie elementy, jak jasność i dobór słów. Możliwość zastosowania,

Tabela 1. Klasyfikacja wyników DISCERN i odpowiadające im poziomy jakości

Wynik DISCERN	Na 100 (%)	Poziom jakości
64–80	80 i powyżej	Doskonały
52–63	65–79	Dobry
41–51	51–64	Przeciętny
30–40	37–50	Słaby
16–29	20–36	Bardzo słabe

Odnosząc się do łatwości, z jaką pacjenci mogli zidentyfikować kolejne kroki, które można podjąć, zmierzono to za pomocą pięciu ukierunkowanych pytań typu „tak/nie". Materiały, które uzyskały wynik 70% lub wyższy, zostały sklasyfikowane jako „zrozumiałe" lub „możliwe do zastosowania" [31, 32]. Po przekroczeniu tego progu wyniki PEMAT stanowiły podstawę do oceny porównawczej. Na koniec dokładność wygenerowanych PEM została oceniona za pomocą skali Likerta dla dezinformacji, w zakresie od 1 do 5. Wynik 1 oznaczał „brak dezinformacji", 3 oznaczało „umiarkowaną dezinformację", a 5 oznaczało „wysoką dezinformację" [27]. Ta dokładna ocena zapewniła wszechstronną analizę jakości, użyteczności i wiarygodności materiałów edukacyjnych dotyczących krótkowzroczności u dzieci.

Ocena czytelności materiałów edukacyjnych

Aby ocenić, jak łatwo jest przeczytać i zrozumieć wszystkie PEM, wykorzystaliśmy dwa dobrze znane wskaźniki czytelności — formuły SMOG i FKGL — za pomocą internetowego kalkulatora czytelności Readable.com. Wskaźniki te oceniają poziom wykształcenia wymagany do zrozumienia materiału, przy czym SMOG koncentruje się na oszacowaniu, ile lat nauki potrzebuje dana osoba, aby zrozumieć dany tekst. SMOG koncentruje się na liczbie wielosylabowych słów w określonej liczbie zdań, zapewniając wynik na poziomie klasy [33]. Wskaźnik FKGL określa czytelność, analizując średnią liczbę sylab w każdym słowie i liczbę słów w każdym zdaniu, dając wynik na poziomie klasy amerykańskiej szkoły [34]. Wyniki w zakresie od 1 do

Wynik 12 odzwierciedla poziom trudności czytania odpowiadający klasom od 1 do 12, wyniki od 13 do 16 wskazują na złożoność tekstu na poziomie akademickim, a każdy wynik 17 lub wyższy oznacza, że tekst wymaga bardziej zaawansowanego zrozumienia niż zazwyczaj zapewnia edukacja akademicka [35, 36]. Poniższe wzory przedstawiają szczegóły tych dwóch wskaźników:

− Flescha Poziom Kincaida

= 0,39 całkowita liczba słów/całkowita liczba zdań

+ 11,8 całkowita liczba sylab/całkowita liczba słów 15,59

$$\text{Poziom SMOG} = 1{,}0430 \times \frac{\text{Liczba wielosylabowych słów}}{\text{Liczba zdań}} \times 30 + 3{,}1291$$

Analiza statystyczna

W celu porównania wskaźników czytelności zastosowaliśmy testy *t* dla dwóch prób, a w przypadku jakości, zrozumiałości, wykonalności i dokładności – testy *U* Manna-Whitneya. Następnie porównaliśmy wyniki czytelności we wszystkich trzech modelach językowych, stosując jednoczynnikową analizę wariancji (ANOVA). Aby zidentyfikować wszelkie istotne różnice w wydajności między modelami, przeprowadziliśmy test Tukey'ego (HSD). Dla wszystkich analiz ustalono poziom istotności p< 0,05. Analizy statystyczne przeprowadzono przy użyciu oprogramowania SPSS (wersja 29, IBM Corp, USA).

WYNIKI

Wszystkie generowane PEM w podpowiedzi A przez ChatGPT-4o (wersja 01-Preview) i ChatGPT-3.5 były dobrej jakości (mediana wyniku DISCERN wyniosła odpowiednio 52,8 i 52,7). Jakość plików PEM wygenerowanych przez ChatGPT-4o (wersja o1 Preview) i ChatGPT-3.5 wykazała znaczny spadek odpowiednio od polecenia A do polecenia B (p= 0,001 i p= 0,013). PEM wygenerowane przez Google Gemini były dość dobrej jakości, a mediana wyniku DISCERN wyniosła 43. Jednak jakość PEM wzrosła od polecenia A do polecenia B (p= 0,02). Wszystkie odpowiedzi wygenerowane przez trzy modele LLM przekroczyły próg 70% wymagany do zaklasyfikowania ich jako „zrozumiałe" (na podstawie PEMAT skala zrozumiałości, która wynosi od 0 do 100%, przy czym wyniki 70% lub wyższe są uważane za „zrozumiałe") (Tabela 2).

W związku z tym, w oparciu o skalę zrozumiałości PAMET, wszystkie wygenerowane PEM zawierały zrozumiałe szczegóły i informacje, miały jasny cel i zawierały tylko najważniejsze informacje, unikając wszelkich rozpraszających szczegółów, miały odpowiedni układ i wygląd, logiczną sekwencję, nagłówki oddzielające sekcje oraz podsumowanie kluczowych punktów. Żadna z odpowiedzi wygenerowanych przez LLM nie spełniała kryteriów pozwalających na zaklasyfikowanie jej jako „action-

„możliwe do zastosowania" (zgodnie z definicją skali wykonalności PEMAT, gdzie wyniki mieszczą się w przedziale od 0 do 100%, a wyniki 70% lub wyższe uznaje się za możliwe do zastosowania). We wszystkich 120 odpowiedziach modele LLM uzyskały stały wynik 40%. Wskazuje to, że treść i struktura wygenerowanych materiałów edukacyjnych nie zawierały jasnych, szczegółowych wskazówek dla pacjentów dotyczących podjęcia konkretnych działań.

Wszystkie 120 odpowiedzi otrzymało wynik 1 w skali dezinformacji Likerta, co wskazuje, że te trzy modele LLM nie wygenerowały żadnych dezinformacji w 120 nowo wygenerowanych PEM.

Nasza analiza czytelności wykazała, że podpowiedź B generowała bardziej czytelne PEM w porównaniu z podpowiedzią A, co odzwierciedlały niższe wyniki SMOG i FKGL zarówno dla ChatGPT-3.5, jak i ChatGPT-4o (o1 Preview) (*p*< 0,001). Poprawa ta znalazła odzwierciedlenie w kluczowych wskaźnikach czytelności, takich jak zmniejszenie liczby sylab, liczby słów, słów o długości 3+ sylab (słowa złożone) oraz liczby zdań (p< 0,05) (Tabela 3).

Wśród tych modeli LLM zarówno ChatGPT-4o (o1 Preview), jak i ChatGPT-3.5 były w stanie wygenerować materiały edukacyjne na poziomie czytania odpowiadającym szóstej klasie lub niższym w odpowiedzi na polecenie B (wyniki FGKL 6± 0,6; wyniki FGKL 6,2± 0,3), podczas gdy Google Gemini nie był w stanie wygenerować materiałów na tym poziomie (wyniki FGKL; 7± 0,6).

W analizie bezpośredniej ChatGPT-4o (o1 Pre-view) wygenerował PEM bardziej czytelne (niższe wyniki SMOG i FGKL; odpowiednio 5,8± 0,7 i 6± 0,6) niż Google Gemini (*p*< 0,001), chociaż różnica nie była statystycznie istotna.

Tabela 2 Porównanie podpowiedzi A i podpowiedzi B: ocena jakości i zrozumiałości dużych modeli językowych w generowanych materiałach edukacyjnych dla pacjentów

LLM	Punkty rozróżniające	Mediana (zakres)	*N*	Średnie rangi	Suma rang	Wartość *p**
Jakość (DISCERN)						
Bliźnięta	Prompt A	43 (43–43)	20	16,52	330	0,021
	Prompt B	49 (37–57)	20	24,51	490	
ChatGPT3.5	Prompt A	52,72 (51–53)	20	24,18	483,52	0,013
	Pytanie B	51 (49–53)	20	16,83	336,54	
ChatGP-4o (01 Podgląd)	Prompt A	52,81 (49–53)	20	25,38	507,53	0,001
	Pytanie B	50 (41–53)	20	15,63	312,52	
Zrozumiałość (PEMAT %)						
Bliźnięta	Prompt A	75 (75–83,3)	20	20,53	410	1,00
	Prompt B	75 (75–83,3)%	20	20,52	410	
ChatGPT3.5	Prompt A	75 (75–83,3)%	20	19	380	0,389
	Prompt B	83,31 (75–83,3)%	20	22	440	
ChatGP-4o (wersja zapoznawcza 01)	PromptA	75 (75–83,3)%	20	20	400	0,771
	Prompt B	83,32 (75–83,3)%	20	21	420	

LLM – duży model językowy, *PEMAT* – narzędzie do oceny materiałów edukacyjnych dla pacjentów

*Test *U* Manna-Whitneya przeprowadzony między promptem A i B (istotność przy p< 0,05)

istotne w porównaniu z ChatGPT-3.5 (p= 0,846) (rys. 1).

Korzystając z podpowiedzi C, trzy modele LLM znacznie poprawiły czytelność istniejących zasobów edukacyjnych online, o czym świadczy wyraźna poprawa średnich wyników czytelności (p≤ 0,001) (rys. 2). Oryginalne materiały edukacyjne online dotyczące krótkowzroczności u dzieci miały średni wynik SMOG wynoszący 10,3± 2,2 i wynik FKGL wynoszący 9,7± 1,9, co znacznie przewyższa zalecany poziom czytania dla szóstej klasy (tabela 4).

Po zastosowaniu modeli LLM zaobserwowano znaczną poprawę (p< 0,001). ChatGPT-3.5 obniżył wyniki czytelności do SMOG na poziomie 7,6± 1,2 i FKGL do 7,7± 1,4. Google Gemini również przyczynił się do poprawy czytelności, osiągając SMOG na poziomie 7,8± 1,3 i FKGL na poziomie 7,5± 1,1, podczas gdy ChatGPT-4o (o1 Preview) wykazał najbardziej znaczącą poprawę, osiągając lub utrzymując się poniżej określonego poziomu czytania dla szóstej klasy (SMOG 5,3± 1,6, FKGL 5,8± 1,5). Ponadto, w bezpośredniej analizie czytelności Ulepszenia wprowadzone przez modele LLM sprawiły, że ChatGPT-4o (o1 Preview) konsekwentnie osiągał znacznie lepsze wyniki niż ChatGPT-3.5 i Google Gemini (p< 0,001 dla obu porównań) (rys. 3). Wyniki te podkreślają transformacyjny potencjał modeli LLM, a w szczególności ChatGPT-4o (o1 Preview), w zakresie upraszczania złożonych materiałów medycznych, czyniąc je bardziej przystępnymi i przyjaznymi dla pacjentów.

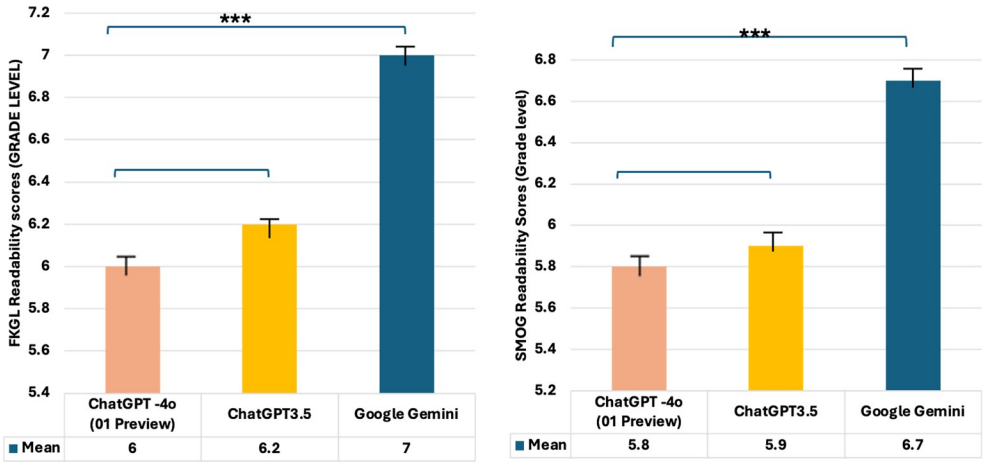
DYSKUSJA

Zgodnie z naszą wiedzą, jest to pierwsze badanie poświęcone analizie sposobu, w jaki modele LLM mogą pomóc rodzicom w zrozumieniu informacji zdrowotnych dostępnych w Internecie oraz w tworzeniu materiałów edukacyjnych dotyczących krótkowzroczności u dzieci. W odróżnieniu od wcześniejszych badań, które opierały się głównie na standardowych testach i powiązanych pytaniach [37–39], nasze badania przyjmują inne podejście, analizując realistyczne scenariusze, w których zaniepokojeni

Tabela 3. Wydajność modeli LLM w oparciu o wskaźniki czytelności dla polecenia A w porównaniu z poleceniem B

Czytelność Wskaźniki	ChatGPT-3.5			ChatGPT-4o (wersja 01 preview)			Google Gemini		
	Prompt A	Prompt B	Wartość *p* (A vs B)	Prompt A	Wskazówka B	Wartość *p* (A vs B)	Wskazówka A	Wskazówka B	Wartość *p* (A vs B)
Sylaby	789,7 (137,5)	562,4 (76,9)	< 0,001	861,9 (145,8)	534,4 (80)	< 0,001	558,9 (109,2)	407,5 (107,8)	< 0,001
Słowa	500,2 (96,9)	381,6 (51,2)	< 0,001	564,5 (91,8)	370,5 (51,6)	< 0,001	326,6 (60,2)	272,2 (60,7)	< 0,001
3+ sylaby słowa	37,3 (6,3)	18,4 (4,1)	< 0,001	36,7 (10,8)	15,35 (5)	< 0,001	35,2 (10,9)	17,6 (7,3)	< 0,001
Zdania	27,5 (5,6)	27,3 (3,7)	< 0,001	31,7 (4,9)	24,5 (4,3)	< 0,001	19,7 (4,2)	17,6 (4,9)	< 0,001
SMOG	8,03 (0,6)	5,9 (0,4)	< 0,001	7,7 (0,7)	5,8 (0,7)	< 0,001	9,07 (0,62)	6,7 (0,8)	< 0,001
Wynik czytelności Flesch-Kin-	7,7 (0,5)	6,2 (0,3)	< 0,001	7,5 (0,58)	6 (0,6)	< 0,001	8,5 (0,7)	7,0 (0,6)	< 0,001
Ocena caid Poziom									

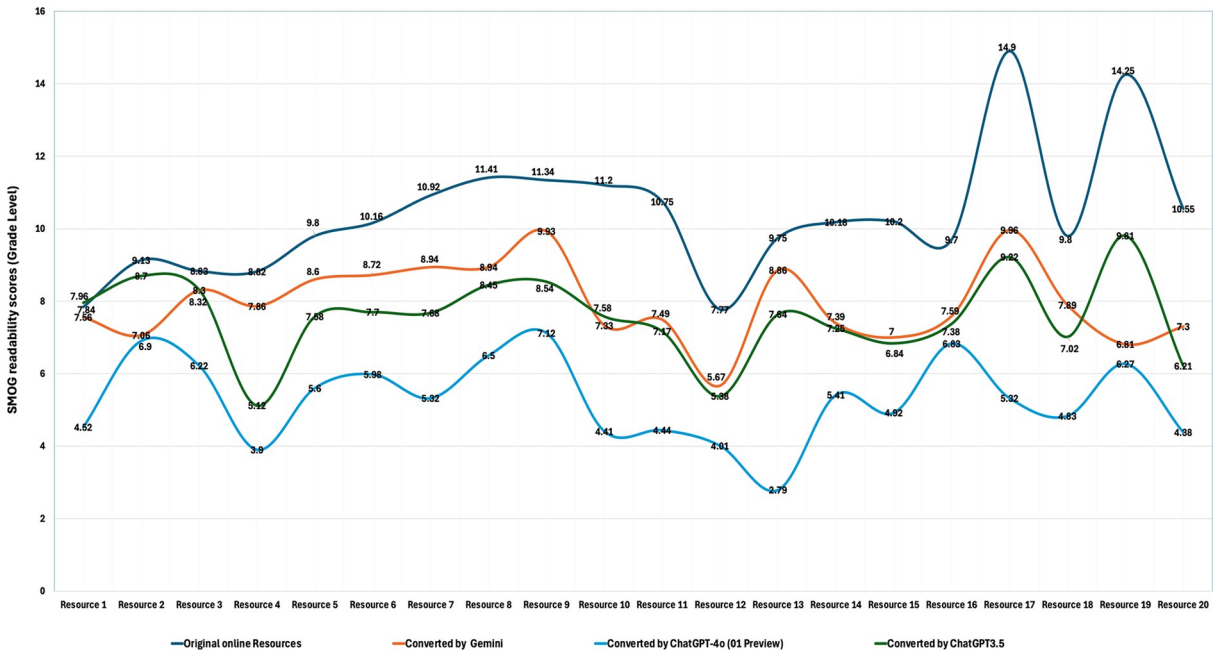
LLM duży model językowy, *SMOG* prosta miara bełkotliwości



Rys. 1 Porównanie wydajności dużych modeli językowych dla promptu B w oparciu o SMOG (prosta miara niezrozumiałego języka) i FKGL (poziom Flesch-Kincaid)

. Jednokierunkowa ANOVA (jednokierunkowa analiza wariancji), test post hoc Tukey'ego

rodzice mogą zwrócić się o pomoc do tych nowych narzędzi. Podkreśla to pilną potrzebę oceny dokładności i wiarygodności odpowiedzi udzielanych przez chatboty LLM w rzeczywistych sytuacjach. W związku z tym oceniliśmy wydajność modeli LLM, a mianowicie ChatGPT-4o (01Preview), ChatGPT-3.5 i Google Gemini, stosując kompleksowe podejście, które obejmowało wyniki DISCERN w celu oceny jakości materiałów, wyniki PEMAT w celu pomiaru ich zrozumiałości i wykonalności, skalę dezinformacji w celu sprawdzenia dokładności oraz wskaźniki czytelności w celu zapewnienia

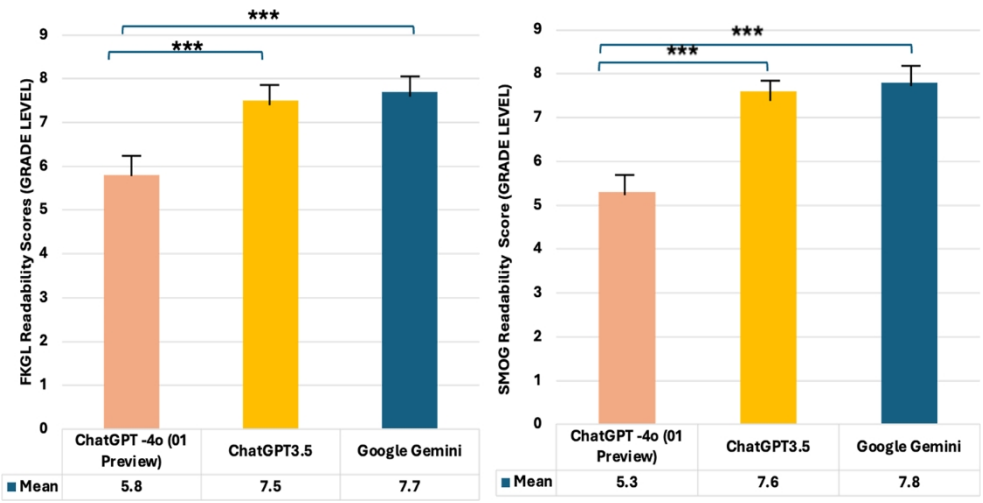


Rys. 2 Czytelność zasobów edukacyjnych online oraz wydajność ChatGPT-4o (01 Preview), ChatGPT-3.5 i Google Gemini w poprawianiu czytelności oryginalnych zasobów online. *SMOG* Prosta miara niezrozumiałego języka

Tabela 4 Czytelność oryginalnych zasobów internetowych i wydajność modeli LLM w zakresie poprawy czytelności oryginalnych zasobów internetowych

Wskaźniki czytelności	Oryginalne zasoby	ChatGPT-4o (01 preview)	Wartość *p* (oryg. vs GPT-4o)	ChatGPT-3.5	Wartość *p* (oryg. vs GPT-3.5)	Google Gemini	Wartość *p* (oryg. vs Gemini)
Sylaby	1457,2 (736,9)	330,4 (159,6)	< 0,001	984,5 (412,7)	0,008	808,4 (205,7)	< 0,001
Słowa	871,0 (392,7)	230,4 (107,9)	< 0,001	649,2 (256,2)	0,21	521,9 (122,4)	< 0,001
Słowa składające się z 3 lub więcej sylab	91,0 (60,7)	13,3 (10,4)	< 0,00	43,0 (23,1)	0,001	38,5 (12,6)	< 0,001
Wyroki	41,5 (17,1)	18,4 (7,8)	0,04	41,0 (16,6)	0,9	31,1 (9,0)	0,02
SMOG Odczyt	10,3 (2,2)	5,3 (1,6)	< 0,001	7,6 (1,2)	< 0,001	7,8 (1,3)	< 0,001
Wynik zdolności Flesch-Kin- Ocena caid Poziom	9,7 (1,9)	5,8 (1,5)	< 0,001	7,7 (1,4)	< 0,001	7,5 (1,1)	< 0,001

LLM duży model językowy, *SMOG* prosta miara bełkotliwości



Rys. 3 Wydajność dużych modeli językowych w zakresie przepisywania oryginalnych materiałów informacyjnych w oparciu o wyniki SMOG (Simple Measure of Gobbledygook) i FKGL (Flesch-Kincaid Grade Level)

. Jednoczynnikowa ANOVA (jednoczynnikowa analiza wariancji), test post hoc Tukey'ego

treść była łatwa do zrozumienia. Wszystkie generowane PEM w poleceniu A przez ChatGPT-4o (o1 Pre-view) i ChatGPT-3.5 były dobrej jakości, podczas gdy PEM generowane przez Google Gemini były średniej jakości. Nie należy tego jednak rozumieć jako wyraźnej przewagi ChatGPT nad Google Gemini. Poprzednie badania wykazały, że każdy model sztucznej inteligencji ma swoje mocne i słabe strony, przy czym Google Gemini jest lepszy w takich aspektach, jak rozumienie języka, ale ma braki w innych aspektach, w których ChatGPT osiąga lepsze wyniki [40]. Wszystkie odpowiedzi wygenerowane przez trzy modele LLM były zrozumiałe, ale żadna z nich nie była praktyczna, ponieważ nie zawierały one jasnych wskazówek krok po kroku dla pacjentów, aby podjęli konkretne działania. Zidentyfikowaliśmy niedociągnięcia w dwóch kluczowych obszarach praktyczności (pozycje 23 i 26 PEMAT), wskazując możliwości poprawy. Po pierwsze, w odpowiedziach brakowało praktycznych narzędzi, takich jak listy kontrolne przypominające o działaniach lub plany działania, które pomogłyby pacjentom w realizacji informacji. Po drugie, ponieważ odpowiedzi były całkowicie oparte na tekście, brakowało im dodatkowej przejrzystości i zaangażowania, które mogłyby zapewnić instruktażowe pomoce wizualne. Warto rozważyć, że włączenie narzędzi do działania dla pacjentów do sformułowań podpowiedzi może być obiecującym obszarem przyszłych badań w celu zwiększenia wykonalności przyszłych PEM.

To odkrycie jest bardzo interesujące, ponieważ pomimo obiecujących wyników sztucznej inteligencji w opracowywaniu planów leczenia pacjentów w poprzednich badaniach [41], nasza praca pokazuje, że nie jest ona wiarygodna w generowaniu planów działania dla pacjentów, którzy chcą uporządkować swoje listy kontrolne w przypadku krótkowzroczności dziecięcej. Najprawdopodobniej nie jest to nieodłączna słabość samej sztucznej inteligencji, ale raczej zbiorów danych, na których została ona wytrenowana [41].

Materiały edukacyjne dotyczące zdrowia muszą być zrozumiałe i łatwe do przeczytania, ale muszą też zawierać informacje wysokiej jakości. W 120 nowo wygenerowanych PEM nie było żadnych błędnych informacji. Nasze wyniki pokazują, że LLM oceniane w tym badaniu nie generowały żadnych „konfabulacji" AI ani fałszywych lub wprowadzających w błąd informacji, co jest częstym problemem zgłaszanym w innych badaniach [42, 43].

Po ocenie 20 najlepszych wyników wyszukiwania, które stanowią ponad 95% całkowitego ruchu internetowego wśród użytkowników internetowych wyszukiwarek [20], wszystkie wyniki przekraczają poziom czytania zalecany przez NIH i AMA dla uczniów szóstej klasy, który ma zapewnić przeciętnemu Amerykaninowi zrozumienie informacji dotyczących zdrowia [6, 8, 44]. Wszystkie ocenione modele LLM wykazały zdolność do przepisywania lub poprawiania czytelności internetowych zasobów edukacyjnych. Jednakże

ChatGPT-4o (o1 Preview) przewyższył inne modele, w tym ChatGPT-3.5 i Google Gemini, znacznie poprawiając czytelność i konsekwentnie spełniając lub pozostając poniżej poziomu zalecanego przez organizacje takie jak NIH i AMA dla przeciętnego Amerykanina w zakresie zrozumienia informacji dotyczących zdrowia [16, 44]. Chat-GPT-4o (o1 Preview) dostosował treść do poziomu czytania odpowiadającego średniej biegłości przeciętnego użytkownika lub poniżej tego poziomu, dzięki czemu informacje stały się bardziej dostępne i łatwiejsze do zrozumienia dla szerszego grona odbiorców. Chociaż ChatGPT-3.5 i Google Gemini poprawiły czytelność internetowych PEM, nie spełniały one konsekwentnie zalecanych standardów ustalonych przez NIH i AMA dotyczących zrozumienia informacji zdrowotnych. Podkreśla to przewagę ChatGPT-4o (o1 Preview) w zakresie upraszczania złożonego języka medycznego do formatu bardziej przystępnego dla użytkowników.

ChatGPT-4o (o1 Preview) wykazał się również przewagą w generowaniu najbardziej czytelnych modeli LLM dotyczących innych chorób. W jednym z badań sprawdzono przydatność ChatGPT-4o (o1 Preview) w nauczaniu osób o schorzeniach dermatologicznych i stwierdzono, że może on tworzyć akapity zrozumiałe dla osób o poziomie czytania od licealnego do wczesnego studenckiego [45]. W innej ocenie zbadano zdolność ChatGPT-4o (o1 Preview) do generowania informacji na temat zapalenia błony naczyniowej oka, prosząc go o napisanie szczegółowych informacji zdrowotnych dotyczących tej choroby, skierowanych do pacjentów. Następnie oceniono adekwatność i czytelność, przy czym tę ostatnią zmierzono za pomocą wzoru FKGL; we wszystkich próbach treść okazała się zarówno odpowiednia, jak i łatwa do dostosowania [46]. W innym bardzo istotnym badaniu naukowcy ocenili wydajność chatbotów LLM w tworzeniu materiałów zdrowotnych skierowanych do pacjentów na temat jaskry dziecięcej, stosując metodę podpowiadania w celu oceny zrozumiałości tekstu. ChatGPT-4o (o1 Preview) wykazał obiecujące wyniki [7]. Może to wynikać z faktu, że ChatGPT-4o (o1 Preview) został przeszkolony na większym zbiorze danych materiałów pisemnych i ma więcej parametrów niż jego wcześniejsze wersje [47]. Jednak obecnie ChatGPT-4o (o1 Preview) jest dostępny tylko w ramach płatnej subskrypcji, co może sprawiać, że jest mniej dostępny dla ogółu społeczeństwa. Nasze ustalenia pokazują, że sposób strukturyzacji polecenia może mieć znaczący wpływ na czytelność odpowiedzi.

wygenerowane przez LLM. Wszystkie trzy modele tworzyły bardziej czytelne treści, gdy otrzymały bardziej szczegółowe zapytanie (prompt B) w porównaniu z mniej szczegółowym (prompt A). Prompt B był bardziej szczegółowy, ponieważ zawierał modyfikator określający docelowy poziom czytelności oraz wyraźną instrukcję dotyczącą zastosowania „formuły czytelności FKGL". Nasze ustalenia podkreślają kluczowe znaczenie stosowania sprawdzonych narzędzi, takich jak formuły czytelności, w celu osiągnięcia pożądanego poziomu czytelności materiałów edukacyjnych. Podejście to jest zgodne z najnowszymi badaniami, które podkreślają, że włączenie takich narzędzi zapewnia, że treść jest zarówno dostępna, jak i skuteczna dla docelowych odbiorców [7, 46].

Badanie to podkreśla ogromny potencjał modeli LLM, w szczególności ChatGPT-4o (o1 Preview), jako narzędzia wypełniającego luki w wiedzy na temat zdrowia i dostarczającego dostosowane do potrzeb, skoncentrowane na pacjencie materiały edukacyjne. Jednak brak praktycznej przydatności generowanych treści podkreśla potrzebę dalszego rozwoju modeli sztucznej inteligencji w celu stworzenia praktycznych, krok po kroku instrukcji dotyczących zdrowia dla użytkowników. Przyszłe badania powinny skupiać się na poszukiwaniu metod optymalizacji wyników modeli LLM, takich jak integracja zestawów danych dotyczących konkretnych dziedzin i udoskonalenie technik inżynierii promptów w celu uzyskania bardziej praktycznych odpowiedzi. Ponadto badania uwzględniające różnorodne konteksty językowe i kulturowe pomogłyby uogólnić te wyniki poza populacje anglojęzyczne i sprostać globalnym wyzwaniom związanym z wiedzą na temat zdrowia. Biorąc pod uwagę rosnące znaczenie elementów wizualnych w komunikacji dotyczącej zdrowia, przyszłe badania powinny również oceniać zdolność modeli LLM do generowania lub rekomendowania wysokiej jakości pomocy wizualnych, takich jak infografiki i narzędzia interaktywne, które uzupełniają materiały tekstowe. Takie postępy mogłyby ułatwić integrację modeli LLM z procesami klinicznymi, umożliwiając pracownikom służby zdrowia tworzenie w czasie rzeczywistym bardziej spersonalizowanych i zrozumiałych zasobów edukacyjnych dla pacjentów.

Podsumowując, wyniki te podkreślają zdolność modeli LLM do pełnienia funkcji wszechstronnego narzędzia służącego do dostarczania jasnych, wysokiej jakości materiałów edukacyjnych dla pacjentów, poprawiającego czytelność internetowych zasobów edukacyjnych i dostępność informacji zdrowotnych, co po dalszym udoskonaleniu i dopracowaniu może otworzyć drzwi do włączenia chatbotów opartych na modelach LLM do zarządzania opieką nad osobami z krótkowzrocznością.

Nasze badanie ma pewne ograniczenia. Ponieważ modele LLM nie są idealne, generowane przez nie odpowiedzi są również ograniczone do danych, na podstawie których zostały wytrenowane. Ponadto sposób formułowania poleceń jest z natury subiektywny. Aby zminimalizować potencjalną stronniczość, starannie podzieliliśmy polecenia na odrębne podzdania kontrolne i modyfikujące oraz przeprowadziliśmy wiele powtarzanych prób dla każdego polecenia, opierając się jednocześnie na ustalonych zasadach inżynierii poleceń [47]. Ponadto ocenialiśmy tylko PEM w języku angielskim, który, choć jest najczęściej używany, nie zawsze jest językiem, w którym pacjenci chcieliby czytać o swoich schorzeniach. Co więcej, oceniając istniejące zasoby internetowe, zdajemy sobie sprawę, że ograniczenie naszego zakresu do 20 najlepszych PEM w Google może potencjalnie wykluczyć strony o niższej pozycji w rankingu, które również mogą dostarczać cennych lub alternatywnych informacji. Jednak uwzględnienie najczęściej odwiedzanych stron internetowych odzwierciedla rzeczywiste wzorce użytkowania i dlatego pozostaje praktyczną i znaczącą próbą. Ponadto niniejsze badanie koncentruje się głównie na populacji Stanów Zjednoczonych, ponieważ większość badań dotyczących czytelności i umiejętności korzystania z informacji zdrowotnych pochodzi z badań przeprowadzonych w USA. Należy jednak pamiętać, że niewystarczająca umiejętność korzystania z informacji zdrowotnych jest powszechnym problemem również w wielu innych dużych krajach [48, 49]. Wreszcie, wiele zasobów dotyczących zdrowia skierowanych do pacjentów zawiera elementy wizualne, takie jak grafiki, obrazy, filmy i demonstracje, które pomagają w lepszym zrozumieniu treści. Jednak w naszym badaniu nie oceniano tych funkcji.

WNIOSKI

W niniejszym badaniu podkreślono obiecujący potencjał dużych modeli językowych, w szczególności Chat-GPT-4o (o1 Preview), który w momencie przeprowadzania badania osiągał najlepsze wyniki spośród ocenianych modeli LLM pod względem generowania wysokiej jakości, czytelnych i zrozumiałych materiałów edukacyjnych dla pacjentów dotyczących krótkowzroczności u dzieci. Chociaż ChatGPT-4o (o1 Preview) przewyższał inne modele pod względem poprawy czytelności i dostępności, istotnym ograniczeniem był brak praktycznych wskazówek w generowanych treściach . Podkreśla to potrzebę ciągłego ulepszania modeli sztucznej inteligencji w celu opracowania bardziej praktycznych informacji zdrowotnych. Przyszłe badania powinny skupiać się na udoskonaleniu inżynierii promptów, uwzględnieniu różnorodnych kontekstów językowych i kulturowych oraz zbadaniu funkcji multimodalnych, takich jak elementy wizualne, aby w pełni wykorzystać potencjał modeli LLM w edukacji pacjentów. Dzięki ciągłym ulepszeniom modele LLM mogą odegrać kluczową rolę w zwiększeniu dostępności, spersonalizowaniu i skuteczności informacji dotyczących opieki zdrowotnej.

Pomoc w zakresie pisania tekstów medycznych/redagowania. Podczas przygotowywania niniejszego manuskryptu autorzy wykorzystali ChatGPT-4 firmy OpenAI do edycji i korekty w celu poprawy czytelności. Następnie autorzy dokładnie przejrzeli i udoskonalili treść w razie potrzeby i ponoszą pełną odpowiedzialność za ostateczną wersję publikacji.

Wkład autorów. Mohammad Delsoz, Amr Hassan, Amin Nabavi, Amir Rahdar, Brian Fowler, Natalie C. Kerr, Lauren Claire Ditta, Mary
E. Hoehn, Margaret M DeAngelis, Andrzej Grzybowski, Yih-Chung Tham i Siamak Yousefi przyczynili się do opracowania koncepcji badania. Mohammad Delsoz, Amr Hassan, Brian Fowler, Lauren Claire Ditta, Margaret M DeAngelis, Andrzej Grzybowski, Yih-Chung Tham i Siamak Yousefi przyczynili się do opracowania projektu badania. Mohammad Delsoz, Amr. Hassan, Amin Nabavi, Amir Rah-dar i Siamak Yousefi przyczynili się do pozyskania danych. Mohammad Delsoz i Siamak Yousefi przyczynili się do analizy i interpretacji danych, wykresów i tabel. Mohammad Delsoz, Amr. Hassan i Siamak Yousefi uzyskali dostęp do każdego zestawu danych i zweryfikowali je w trakcie badania. Planowanie i realizacja badań były nadzorowane przez Siamaka Yousefi, Yih-Chunga Thama, Margaret M DeAngelis, Andrzeja Grzybowskiego, Briana Fowlera i Lauren Claire Ditta. Mohammad Delsoz, Amr Hassan, Lauren Claire Ditta, Andrzej Grzybowski, Yih-Chung Tham i Siamak Yousefi przygotowali szkic manuskryptu. Wszyscy autorzy przeczytali i zatwierdzili ostateczną wersję manuskryptu.

Finansowanie. Praca została sfinansowana przez Hamilton Eye Institute (SY). Fundator nie miał wpływu na projekt badania, gromadzenie i analizę danych, decyzję o publikacji ani przygotowanie manuskryptu. Opłata za szybką obsługę czasopisma została sfinansowana przez dr Siamaka Yousefi.

Dostępność danych. Wszystkie dane niezbędne do powtórzenia naszych wyników znajdują się w pliku uzupełniającym, z wyjątkiem surowych ocen graderów, które są dostępne na żądanie.

Oświadczenia

Konflikt interesów. Mohammad Delsoz, Amr Hassan, Amin Nabavi, Amir Rahdar, Brian Fowler, Natalie C. Kerr, Lauren Claire Ditta, Mary E. Hoehn, Margaret M DeAngelis i Yih-Chung Tham nie mają nic do ujawnienia. Andrzej Grzybowski jest członkiem komitetu redakcyjnego czasopisma *„Ophthalmology and Therapy"*. Andrzej Grzybowski nie brał udziału w wyborze recenzentów manuskryptu ani w żadnej z późniejszych decyzji redakcyjnych. Siamak Yousefi: Otrzymał prototypowe urządzenia od firm Remidio, M&S Technologies oraz Visrtucal Fields. Świadczy usługi konsultacyjne dla firm InsightAEye i Enolink.

Zgoda komisji etycznej. Badanie zostało zwolnione z oceny komisji etycznej Centrum Nauk o Zdrowiu Uniwersytetu Tennessee, ponieważ nie obejmowało ono uczestników ani ich danych osobowych, a skupiało się na ocenie działania najnowszych modeli sztucznej inteligencji. Skupienie się na publicznie dostępnych danych i tekstach generowanych przez sztuczną inteligencję zapewniło zgodność z normami prywatności i etyki badawczej. Badanie odbyło się w okresie od października do grudnia 2024 r., zgodnie z zasadami Deklaracji Helsińskiej.

BIBLIOGRAFIA

1. Liang J, Pu Y, Chen J, et al. Global prevalence, trend and projection of myopia in children and adolescents from 1990 to 2050: a comprehensive systematic review and meta-analysis. Br J Ophthalmol. 2024. https:// doi. org/ 10. 1136/ bjo-2024-325427.

2. Modjtahedi BS, Ferris FL, Hunter DG, Fong DS. Public health burden and potential interventions for myopia. Ophthalmology. 2018;125(5):628–30.

3. Schweitzer K. W związku z rosnącą liczbą przypadków krótkowzroczności u dzieci eksperci zalecają spędzanie czasu na świeżym powietrzu i uznanie tej wady wzroku za chorobę. JAMA. 2024;332(19):1599–601. https://doi.org/10.1001/jama.2024.21043.

4. Morgan IG, Wu P-C, Ostrin LA i in. Czynniki ryzyka krótkowzroczności IMI. Investig Ophthalmol Vis Sci. 2021;62(5):3–3. https:// doi. org/ 10. 1167/ iovs. 62.5.3.

5. Huang J, Wen D, Wang Q, et al. Porównanie skuteczności 16 interwencji w zakresie kontroli krótkowzroczności u dzieci: metaanaliza sieciowa. Ophthalmology. 2016;123(4):697–708.

6. Stossel LM, Segar N, Gliatto P, Fallar R, Karani R. Czytelność materiałów edukacyjnych dla pacjentów dostępnych w placówkach opieki zdrowotnej. J Gen Intern Med. 2012;27:1165–70.

7. Dihan Q, Chauhan MZ, Eleiwa TK, et al. Wykorzystanie dużych modeli językowych do tworzenia materiałów edukacyjnych na temat jaskry dziecięcej. Am J Ophthal-mol. 2024;265:28–38.

8. Rooney MK, Santiago G, Perni S, et al. Czytelność materiałów edukacyjnych dla pacjentów pochodzących z wpływowych czasopism medycznych: analiza 20 lat. J Patient Exp. 2021;8:2374373521998847.

9. Raja H, Huang X, Delsoz M, et al. Diagnozowanie jaskry na podstawie zbioru danych z badania leczenia nadciśnienia ocznego przy użyciu generatywnego, wstępnie wytrenowanego transformatora czatu

jako dużego modelu językowego. Ophthal-mol Sci. 2025;5(1):100599.

10. Huang X, Raja H, Madadi Y, et al. Przewidywanie jaskry przed wystąpieniem za pomocą chatbota opartego na dużym modelu językowym. Am J Ophthalmol. 2024. https://doi.org/ 10.1016/j.ajo.2024.06.035.

11. Delsoz M, Madadi Y, Raja H, et al. Skuteczność ChatGPT w diagnostyce chorób rogówki. Cornea. 2024;43(5):664–70.

12. Delsoz M, Raja H, Madadi Y, et al. Wykorzystanie ChatGPT do wspomagania diagnostyki jaskry na podstawie opisów przypadków klinicznych. Ophthalmol Ther. 2023;12(6):3121–32.

13. Madadi Y, Delsoz M, Lao PA, et al. ChatGPT wspomagający diagnostykę chorób neurookulistycznych na podstawie opisów przypadków. J Neuroophthalmol. 2022;128:1356.

14. Madadi Y, Delsoz M, Khouri AS, Boland M, Grzybowski A, Yousefi S. Zastosowania robotów i chatbotów opartych na sztucznej inteligencji w okulistyce: najnowsze osiągnięcia i przyszłe trendy. Curr Opin Ophthalmol. 2024;35(3):238–43.

15. Bellanda VC, Santos MLD, Ferraz DA, Jorge R, Melo GB. Zastosowania ChatGPT w diagnostyce, leczeniu, edukacji i badaniach chorób siatkówki: przegląd zakresu badań. Int J Retina Vitreous. 2024;10(1):79.

16. Weiss BD. Wiedza na temat zdrowia. Am Med Assoc. 2003;253:358.

17. System oceny czytelności. Formuły czytelności. https://www.readabilityformulas.com. Dostęp: 7 marca 2024 r.

18. Delsoz M, Raja H, Madadi Y i in. Odpowiedź na: list do redakcji dotyczący „Wykorzystania ChatGPT do pomocy w diagnozowaniu jaskry na podstawie opisów przypadków klinicznych". Ophthalmol Ther. 2024;13(6):1817–

19. OpenAI. Przedstawiamy ChatGPT. OpenAI. 2022.

20. Insights C. Wartość pozycjonowania wyników wyszukiwania Google. Westborough: Chitika; 2013. str. 0–10.

21. Morahan-Martin JM. Jak użytkownicy Internetu znajdują, oceniają i wykorzystują informacje zdrowotne online: przegląd międzykulturowy. CyberPsychol Behav. 2004;7(5):497–510. https://doi.org/10.1089/cpb. 2004.7.497.

22. System oceny czytelności. Formuły czytelności. https://www.readabilityformulas.com/.

23. Martin CA, Khan S, Lee R, et al. Czytelność i przydatność internetowych materiałów edukacyjnych dla pacjentów cierpiących na jaskrę. Ophthalmol Glaucoma. 2022;5(5):525–30.

24. Decker H, Trang K, Ramirez J, et al. Chatbot oparty na dużym modelu językowym a dokumentacja świadomej zgody generowana przez chirurga dla typowych procedur. JAMA Netw Open. 2023;6(10):e2336997.

25. Crabtree L, Lee E. Ocena czytelności i jakości internetowych materiałów edukacyjnych dla pacjentów dotyczących leczenia jaskry z otwartym kątem przesączania. BMJ Open Ophthalmol. 2022;7(1):e000966.

26. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: narzędzie do oceny jakości pisemnych informacji zdrowotnych dla konsumentów dotyczących wyborów terapeutycznych. J Epidemiol Community Health. 1999;53(2):105–11.

27. Pan A, Musheyev D, Bockelman D, Loeb S, Kabar-riti AE. Ocena odpowiedzi chatbota opartego na sztucznej inteligencji na najczęściej wyszukiwane pytania dotyczące raka. JAMA Oncol. 2023;9(10):1437–40.

28. San Giorgi MRM, de Groot OSD, Dikkers FG. Ocena jakości i czytelności stron internetowych związanych z nawracającym brodawczakiem dróg oddechowych. Laryngoscope. 2017;127(10):2293–7. https://doi.org/10.1002/lary.26521.

29. Shoemaker SJ, Wolf MS, Brach C. Narzędzie do oceny materiałów edukacyjnych dla pacjentów (PEMAT) i instrukcja obsługi. Rockville: Agencja ds. Badań i Jakości Opieki Zdrowotnej; 2020.

30. Shoemaker SJ, Wolf MS, Brach C. Opracowanie narzędzia do oceny materiałów edukacyjnych dla pacjentów (PEMAT): nowy sposób pomiaru zrozumiałości i przydatności drukowanych i audiowizualnych informacji dla pacjentów. Patient Educ Couns. 2014;96(3):395–403.

31. Veeramani A, Johnson AR, Lee BT, Dowlatshahi AS. Czytelność, zrozumiałość, użyteczność i wrażliwość kulturowa internetowych materiałów edukacyjnych dla pacjentów (PEMS) dotyczących rekonstrukcji kończyn dolnych: badanie przekrojowe. Plast Surg. 2024;32(3):452–9.

32. Loeb S, Sengupta S, Butaney M, et al. Rozpowszechnianie błędnych i tendencyjnych informacji na temat raka prostaty w serwisie YouTube. Eur Urol. 2019;75(4):564–7.

33. Edmunds MR, Barry RJ, Denniston AK. Ocena czytelności internetowych informacji dla pacjentów okulistycznych. JAMA Ophthalmol. 2013;131(12):1610–6. https://doi.org/10.1001/jamaophthalmol.2013. 5521.

34. Lois C, Edward L. Ocena czytelności i jakości internetowych materiałów edukacyjnych dla pacjentów dotyczących leczenia jaskry z otwartym kątem przesączania. BMJ Open Ophthalmol. 2022;7(1):e000966. https://doi.org/10.1136/bmjophth-2021-000966.

35. Kincaid P, Fishburne RP, Rogers RL, Chissom BS. Wyprowadzenie nowych wzorów na czytelność (automatyczny wskaźnik czytelności, liczba mgły i wzór Flescha na łatwość czytania) dla personelu marynarki wojennej. Mil-lington: Naval Air Station Memphis; 1975.

36. Mc Laughlin GH. Klasyfikacja SMOG — nowa formuła czytelności. J Read. 1969;12(8):639–46.

37. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Ocena działania ChatGPT w okulistyce: analiza jego zalet i wad. Ophthalmol Sci. 2023;3(4):100324.

38. Mihalache A, Popovic MM, Muni RH. Wydajność chatbota opartego na sztucznej inteligencji w ocenie wiedzy okulistycznej. JAMA Ophthalmol. 2023;141(6):589–97.

39. Lim ZW, Pushpanathan K, Yew SME i in. Benchmarking wydajności dużych modeli językowych w leczeniu krótkowzroczności: analiza porównawcza ChatGPT-3.5, ChatGPT-4.0 i Google Bard. EBioMed-icine. 2023;95:104770.

40. Masalkhi M, Ong J, Waisberg E, Lee AG. Sztuczna inteligencja Gemini firmy Google DeepMind a ChatGPT: analiza porównawcza w okulistyce. Eye. 2024;38(8):1412–7. https:// doi. org/ 10. 1038/ s41433-024-02958-w.

41. Satapathy SK, Kunam A, Rashme R, Sudarsanam PP, Gupta A, Kumar HK. Planowanie leczenia wspomagane sztuczną inteligencją w przypadku wszczepiania implantów dentystycznych: plany kliniczne a plany generowane przez sztuczną inteligencję. J Pharm Bioallied Sci. 2024;16(Suppl 1):S939.

42. Hua H-U, Kaakour A-H, Rachitskaya A, Srivastava S, Sharma S, Mammo DA. Ocena i porównanie streszczeń naukowych dotyczących okulistyki oraz odniesień przez obecne chatboty oparte na sztucznej inteligencji. JAMA Ophthalmol. 2023;141(9):819–24. https://doi.org/10.1001/jamaophthalmol.2023.3119 .

43. Brender TD. Medycyna w erze sztucznej inteligencji: hej chatbot, napisz mi H&P. JAMA Intern Med. 2023;183(6):507–8. https://doi.org/10.1001/jamainternmed.2023.1832.

44. Doak LG, Doak CC, Meade CD. Strategie poprawy materiałów edukacyjnych dotyczących raka. Oncol Nurs Forum. 1996;23:1305–12.

45. Mondal H, Mondal S, Podder I. Wykorzystanie ChatGPT do pisania artykułów edukacyjnych dla pacjentów cierpiących na choroby dermatologiczne: badanie pilotażowe. Indian Dermatol Online J. 2023;14(4):482–6.

46. Kianian R, Sun D, Crowell EL, Tsui E. Wykorzystanie dużych modeli językowych do generowania materiałów edukacyjnych na temat zapalenia błony naczyniowej oka. Ophthalmol Retina. 2024;8(2):195–201.

47. Brin D, Sorin V, Vaid A, et al. Porównanie wydajności ChatGPT i GPT-4 w ocenie umiejętności miękkich USMLE. Sci Rep. 2023;13(1):16492.

48. DeWalt DA, Berkman ND, Sheridan S, Lohr KN, Pignone MP. Umiejętność czytania i pisania a wyniki zdrowotne: systematyczny przegląd literatury. J Gen Intern Med. 2004;19:1228–39.

49. Schillinger D. Czynniki społeczne, umiejętność czytania i pisania w zakresie zdrowia oraz nierówności: powiązania i kontrowersje. HLRP Health Literacy Res Pract. 2021;5(3):e234–43.