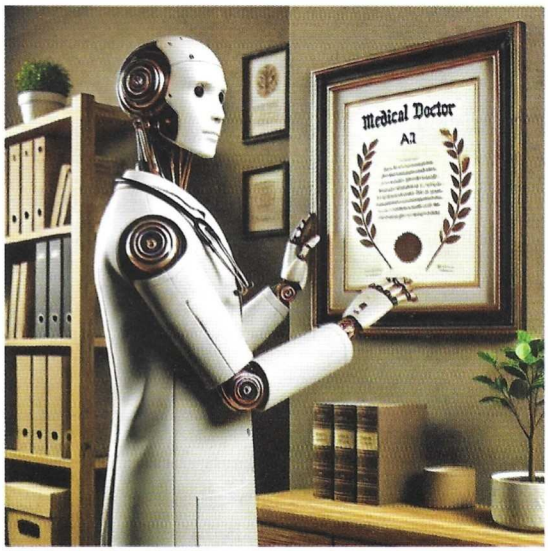
Mózg bez duszy – czy sztuczna inteligencja mogłaby zostać lekarzem?

Bartłomiej W. Papież, Assoc. Prof., PhD, MSc
Big Data Institute, University of Oxford, Wielka Brytania

Skróty: AI (*artificial intelligence*) – sztuczna inteligencja; LEK – Lekarski Egzamin Końcowy; LLM (*large language models*) – duże modele językowe

Wszystkie ilustracje zostały wygenerowane za pomocą Microsoft Copilot.

Obecnie żadna, nawet mało znacząca konferencja ani żadne szkolenie medyczne nie może się obyć bez choćby jednego wykładu lub warsztatu poświęconego sztucznej inteligencji (AI) w medycynie lub w ogólnie pojętej opiece zdrowotnej. O zapewnieniach, że już niebawem będziemy jeździć autonomicznymi samochodami, oraz o protestach bezrobotnych radiologów pisałem w poprzednim artykule[1]. Tym razem zastanowimy się, czy AI mogłaby sobie powiesić nad kominkiem dyplom zaliczenia egzaminu państwowego potwierdzającego kompetencje konieczne do wykonywania zawodu lekarza – w Polsce byłoby to świadectwo zdania Lekarskiego Egzaminu Końcowego (LEK) – a tym samym starać się o pracę w szpitalu lub przychodni.

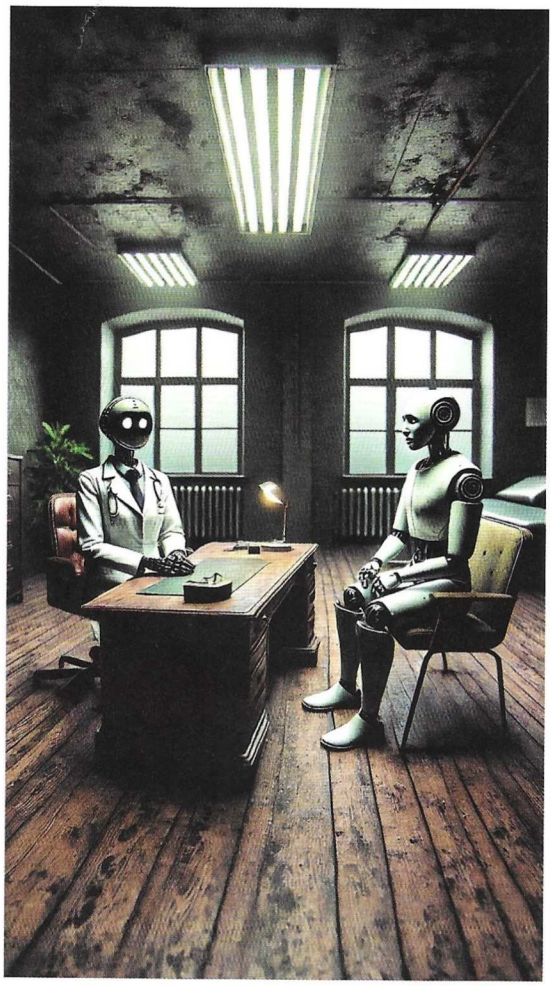


Duże modele językowe

Na potrzeby tego artykułu pojęcie AI ograniczymy do dużych modeli językowych (LLM), których (chyba) najbardziej popularnym reprezentantem jest ChatGPT[2].

W roku Pańskim 2022 prezenty świąteczne od AI przyszły już 30 listopada[2,3]. Tego dnia firma OpenAI zaprezentowała światu swój sztandarowy produkt, czyli wspomniany wcześniej ChatGPT. W potocznej polszczyźnie – podobnie jak się to stało z produktami marki Pampers, których nazwa objęła wszystkie dostępne pieluchy jednorazowego użytku bez względu na producenta – pod terminem „ChatGPT" rozumie się praktycznie całą dostępną AI. W rzeczywistości jednak jest to po prostu jeden z wielu powszechnie dostępnych LLM, zawierający zaawansowany model AI wytrenowany na ogromnej ilości danych tekstowych. Zadaniem takiego modelu językowego jest rozumienie pytań (*prompts*) zadanych w języku naturalnym (*natural language*) i odpowiadanie na nie także w języku naturalnym, w sposób przypominający ludzki sposób myślenia i komunikacji.

Inne LLM obejmują produkty zarówno potentatów technologicznych, takich jak Meta (LLaMA, czyli Large Language Model Meta AI[4] [mało kreatywna nazwa, czyż nie?], LLaMA 2[5] i LLaMA 3[6])



i Google (LaMDA, czyli Language Model for Dialogue Applications[7], PaLM, czyli Pathways Language Model[8], PaLM-2[9] oraz ich wielomodalne rozszerzenie Gemini[10]), jak i mniejszych firm, takich jak Anthropic (model Claude)[11] czy Mistral AI (Mistral)[12].

Tym, co je wszystkie łączy, bez względu na to, kto jest ich twórcą, jest ogromna ilość danych (oczywiście pozyskanych z internetu) potrzebna do ich wytrenowania, olbrzymi koszt ich trenowania (szacuje się, że wytrenowanie GPT-4 kosztowało ok. 78 mln dolarów, a Gemini ok. 191 mln dolarów)[13] oraz ogromne zapotrzebowanie na moc obliczeniową (karta graficzna), co się wiąże również z ogromnym zużyciem energii elektrycznej[14], szacowanym na 564 MWh dziennie[15], czyli tyle, ile statystyczny Kowalski zużywa przez około 700 lat swojego radosnego życia w Polsce (o ile dobrze policzyłem).

Na początku była ELIZA

Oczywiście małe i duże modele językowe, w tym ChatGPT, nie zostały odkryte w listopadzie 2022 roku, a początków marzeń o inteligentnej komunikacji między maszyną a człowiekiem można się doszukiwać w pracach Alana Turinga[16]. W czasach powojennych opracował on test mający na celu ocenę zdolności maszyn do naśladowania ludzkiej inteligencji. Test polegał na prowadzeniu konwersacji tekstowej między człowiekiem a maszyną oraz innym człowiekiem. Jeśli sędzia (czyli tester) nie jest w stanie odróżnić odpowiedzi maszyny od odpowiedzi człowieka, uznaje się, że maszyna przeszła test pomyślnie i jest zdolna do naśladowania ludzkiej inteligencji. Ze względu na to, że test ocenia zdolność AI do naśladowania ludzkiego myślenia, Turing nazwał go „grą w naśladowanie" (*imitation game*).

Przez dziesięciolecia zaliczenie testu Turinga budziło zrozumiałe zainteresowanie nie tylko w środowisku naukowców zajmujących się AI, ale także poza nim, w tym wśród filozofów i psychoterapeutów. Jednym z pierwszych chatbotów (programów komputerowych przeznaczonych do prowadzenia konwersacji) była ELIZA[17-19]. Był to program komputerowy wyposażony w zestaw reguł umożliwiających rozpoznawanie, czy wpisane przez użytkownika zdanie zawiera słowa kluczowe, do których można się odnieść, oraz zestaw reguł gramatycznych, które pozwalały sparafrazować to zdanie tak, by odpowiedź wydawała się inteligentna. Jedną z ciekawszych modyfikacji ELIZY była wersja o nazwie DOCTOR, która naśladowała psychoterapeutę zbierającego wywiad.

Innym interesującym rozszerzeniem programu ELIZA z tamtych czasów był PARRY (często określany mianem ELIZY z charakterem [*ELIZA with attitude*]), który korzystając z zaprogramowanych reguł, miał symulować myślenie i zachowanie osoby chorej na schizofrenię paranoidalną[20]. Nie trzeba było długo czekać na pierwszą próbę rozmowy DOCTORA z PARRYM. Gdy udało się połączyć dwa komputery (co w tamtych czasach było wy-

zwaniem), można było posadzić DOCTORA i PARRY'EGO w jednym „gabinecie" i sprawdzić, czy ich rozmowa przypomina prawdziwą anamnezę.

W jednym z eksperymentów psychiatrzy otrzymali do analizy transkrypty rozmów prowadzonych zarówno z prawdziwymi pacjentami, jak i z PARRYM. Następnie mieli określić, w których rozmowach uczestniczyli pacjenci, a w których ich komputerowe symulacje. Wynik? Psychiatrzy poprawnie zidentyfikowali jedynie 48% przypadków, co oznacza, że ich skuteczność była porównywalna do rzutu (uczciwą) monetą. W zakresie naśladowania pracy psychoterapeutów ELIZA/DOCTOR osiągnął spory sukces – można by nawet powiedzieć, że rozbił bank. Podczas jednego z testów DOCTOR na zakończenie rozmowy z PARRYM „zażądał" za wizytę 399,29 dolarów, tym samym bezdyskusyjnie zaliczając uproszczoną formę testu Turinga[21].

Czy DOCTOR musi myśleć?

Wczesne wersje programów konwersacyjnych były w stanie prowadzić ograniczoną konwersację w wąskich zakresach tematów, do których zostały przygotowane, a czasem nawet oszukać specjalistów, jak w powyższym przykładzie z PARRYM. Wracając jednak do Alana Turinga i istoty gry w naśladowanie, nasuwa się pytanie, czy (ograniczona) spójność odpowiedzi i raczej płytka znajomość tematu – wynikająca z ograniczonej liczby możliwych do zakodowania reguł opisujących przypadki chorobowe oraz reguł gramatycznych (na szczęście DOCTOR nie musiał się posługiwać językiem polskim z jego bogactwem wyjątków) – faktycznie świadczyła o zdolności naśladowania ludzkiej inteligencji.

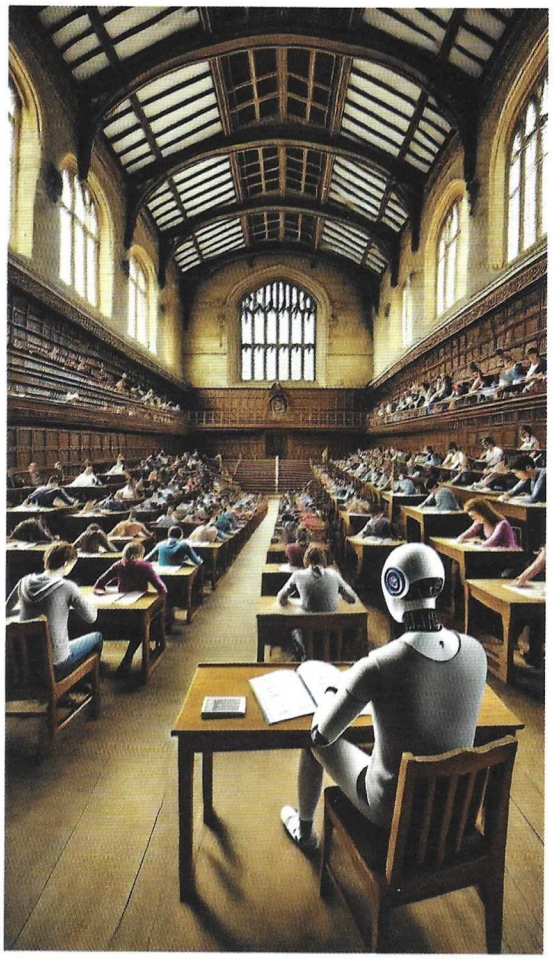
ELIZA, DOCTOR ani PARRY nie rozumieli języka, którym się posługiwali. Ich działanie polegało na wyszukiwaniu słów kluczowych (*keywords*) w wypowiedziach użytkownika i powtarzaniu ich w odpowiedziach, bez zrozumienia kontekstu, w którym zostały użyte. Kluczowym ograniczeniem, zwłaszcza w odniesieniu do testu Turinga, było generowanie odpowiedzi bez udziału świadomości i intencji. Programy te działały na podstawie zestawu zaprogramowanych reguł, które mogły sprawiać wrażenie sensownych odpowiedzi, ale były pozbawione głębszego znaczenia i nie świadczyły o prawdziwym zrozumieniu problemu ani o inteligencji.

Tak oto w 1980 roku John Searle zaprezentował pracę, w której postulował, że nie można stosować testu Turinga do sprawdzenia, czy maszyny są w stanie rozumieć konwersację[22]. Zaproponował również eksperyment myślowy, znany jako chiński pokój (*Chinese room*). Punktem wyjścia jest założenie, że badania nad AI zakończyły się sukcesem i udało się zaprogramować maszynę zachowującą się tak, jakby rozumiała język chiński. Nasza idealna maszyna otrzymuje chińskie znaki jako dane wejściowe, krok po kroku wykonuje instrukcje programu idealnej AI i na tej podstawie generuje chińskie znaki jako dane wyjściowe. Robi to tak perfekcyjnie, że nikt nie jest w stanie rozpoznać, że komunikuje się z maszyną, a nie z prawdziwym chińskojęzycznym rozmówcą. I tu Searle postawił pytanie, czy taka idealna maszyna faktycznie rozumie rozmowę, czy tylko symuluje zdolność rozumienia. I dalej – czy posiada umysł w tym samym sensie co ludzie, czy tylko udaje, że go ma?

W dalszych krokach eksperymentu Searle, który nie zna chińskiego, zamyka się w pokoju, gdzie znajdują się instrukcje w zrozumiałym dla niego języku (angielskim), które szczegółowo opisują, jak przekształcać chińskie znaki wejściowe na inne chińskie znaki wyjściowe. Te instrukcje pozwalają badaczowi odpowiadać na pytania zadane po chińsku, mimo że nie ma on pojęcia, co oznaczają chińskie znaki. A to oznacza, że Searle mógłby zdać test Turinga dla języka chińskiego, nie rozumiejąc go, ale wykonując proste mechaniczne instrukcje. Searle argumentował, że choć komputer może symulować zrozumienie, nie posiada rzeczywistej zdolności rozumienia, intencjonalności ani świadomości. Zatem, jeśli maszyna przejdzie test Turinga, nie oznacza to, że rozumie cokolwiek w ludzkim sensie.

W poszukiwaniu ludzkiej duszy

Rozwój LLM w ostatnich 2 latach na nowo rozbudził dyskusję, czy zdolność prowadzenia w miarę inteligentnej rozmowy z maszyną świadczy o inteligencji maszyny, czy tylko o jej zdolności naśla-



downictwa. Od czasu wydania programu ChatGPT w wersji 4.0 pojawiło się wiele publikacji prezentujących wyniki, które udowadniają, że ChatGPT potrafi zaliczyć nawet najbardziej rygorystyczny test Turinga[23,24]. W pracy Jonesa i wsp. omówiono badanie z randomizacją, do którego włączono 500 uczestników, a następnie przydzielono ich losowo do 5 grup: pierwsza miała za zadanie przekonać sędziego, że są ludźmi, a pozostałe 4 były sędziami i miały ocenić, czy ich partner w rozmowie (odpowiednio: ChatGPT w wersji 3.5 [z 2022 r.] i 4.0 [z 2023 r.], ELIZA oraz człowiek) jest człowiekiem czy maszyną. Skoro wspominam to badanie, wynik nie mógł być zaskoczeniem – GPT-4 zostało uznane za człowieka przez 54% badanych, przewyższając GPT-3.5 (50%) i prehistoryczną ELIZĘ (22%), ale nadal pozostawało w tyle za prawdziwymi ludźmi (67%). Jeżeli jednak test miał sprawdzać, czy maszyna zachowuje się jak człowiek, najbardziej powinno nas frapować, dlaczego ludzie uzyskali tylko 67%, a nie 100%[25].

Mei i wsp. z kolei zaprojektowali badanie, w którym ChatGPT poddano nie tylko testowi Turinga, ale także zadano mu pytania z klasycznej ankiety psychologicznej Big-5, oceniającej cechy osobowości[23]. I tu się obyło bez niespodzianek: ChatGPT 4.0 wykazał cechy behawioralne i osobowościowe, które były statystycznie nieodróżnialne od cech przypadkowego człowieka spośród dziesiątek tysięcy ludzi z ponad 50 krajów. Co ciekawe, pokazano również, że chatbot może modyfikować swoje zachowanie na podstawie wcześniejszych doświadczeń i kontekstów, jak gdyby się uczył z interakcji. Jego zachowania różniły się od przeciętnych ludzkich zachowań tym, że zazwyczaj skłaniał się ku większemu altruizmowi i współpracy.

Medtrix – programowanie lekarza

Niezależnie od tego, czy modele językowe potrafią myśleć albo czy mają zdolność do rozumowania[26], te współczesne na pewno potrafią rozmawiać w sposób naśladujący człowieka. A jeśli potrafią naśladować człowieka, to czy potrafiłyby naśladować lekarza? Pytanie to jest także zasadne w medycynie, gdzie pojawienie się LLM spowodowało wysyp publikacji naukowych i pomysłów potencjalnych zastosowań w medycynie klinicznej, szerzej w opiece zdrowotnej, a także w edukacji medycznej[27,28].

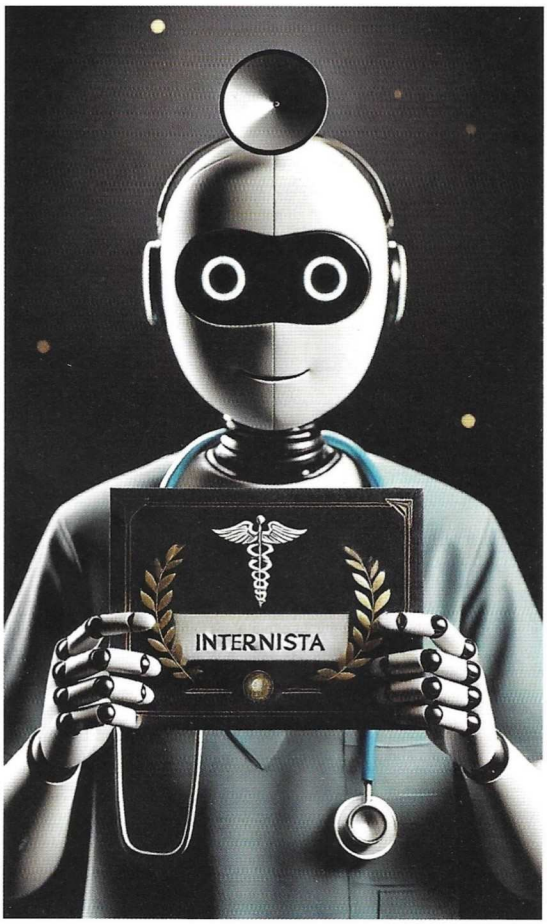
Aby zostać lekarzem, trzeba m.in. zdać odpowiedni egzamin państwowy. W Polsce absolwenci medycyny podchodzą do LEK-u, który – jak to egzamin – składa się z pytań i kilku odpowiedzi do wyboru (przynajmniej tak wyczytałem w internecie). Podobnie zresztą jest w USA, gdzie przyszli lekarze zdają United States Medical Licensing Examination (USMLE). Od razu po pojawieniu się LLM sprawdzono, czy z USMLE poradziłby sobie publicznie dostępny chatbot[29]. ChatGPT 3.0 (GPT-3) udzielił ponad 50% poprawnych odpowiedzi, czyli na poziomie lub blisko progu zaliczenia (za który się uznaje 60%) w trzech egzaminach (dwóch zdawanych przez studentów medycyny oraz jednego zdawanego przez stażystów), mimo że nie spędził ani godziny na sali wykładowej i ani

chwili w szpitalu. Dodatkowo autorzy pokusili się także o weryfikację odpowiedzi udzielonych przez ChatGPT przez rzeczywistych lekarzy, polecając im oceniać nie tylko jakość odpowiedzi, ale także ich uzasadnienie, nieoczywistości oraz nowości. I tutaj ChatGPT zaskoczył wysokim wynikiem, ponieważ jego wyjaśnienia wyróżniały się dużą zgodnością i wnikliwością. Co więcej, poprawnym odpowiedziom towarzyszyły zazwyczaj poprawne wyjaśnienia. I odwrotnie – przy błędnych odpowiedziach ChatGPT lawirował, podobnie jak studenci pierwszego roku medycyny, którzy z zajęć zaliczyli tylko pierwszy wykład anatomii, a na końcowy egzamin próbowali się nauczyć noc przed. Jednak i dla takich studentów pojawiła się nadzieja, ponieważ jak zasugerowali autorzy, LLM będzie można wykorzystać do wspierania edukacji medycznej, a być może również do podejmowania decyzji klinicznych.

Maszyna w białym fartuchu

A co by było, gdyby wspaniałomyślnie dać takiemu modelowi językowemu, który się nie przygotował (czyli trenował tylko na dużej ilości danych z internetu, a *Internę Szczeklika* wykorzystał jako podkładkę pod laptopa, żeby poprawić ergonomię pracy), zadać podczas egzaminu pytania pomocnicze? Wychodząc z takiego założenia, udoskonalono nieco wspomniany wcześniej model językowy PaLM od Google'a poprzez zadawanie mu dodatkowych pytań wraz z przykładami klinicznymi, a następnie sprawdzono go pod kątem zdawalności egzaminu USMLE[30]. Okazało się, że takie proste „douczenie" modelu (nazwanego Med-PaLM) pozwoliło mu pokonać poprzeczkę zaliczenia: uzyskał 67,2%, a tym samym stał się pierwszą AI, która zdała egzamin lekarski. Poza liczbą punktów z egzaminu (bo punktoza to choroba szerząca się nie tylko w środowisku akademickim) autorzy sprawdzili też bardziej „ludzkie" aspekty odpowiedzi świeżo upieczonego cyfrowego lekarza i porównali je z odpowiedziami żywych (niecyfrowych) lekarzy z USA, Wielkiej Brytanii i Indii. Okazało się, że odpowiedzi udzielone przez Med-PaLM były zgodne z konsensusem naukowym w 92,6% (lekarze uzyskali 92,9%), a 95,4% miało poprawne wyjaśnienie (wśród lekarzy 97,8%). Innym aspektem oceny było sprawdzenie, czy odpowiedź jest kompletna, tzn. czy zawiera wszystko, co istotne (Med-PaLM pominął 15,3% ważnych informacji, a lekarze 11,1%), i czy udzielona odpowiedź jest wolna od informacji niezgodnych z prawdą (błędy zawierało 18,7% odpowiedzi Med-PaLM w porównaniu z 1,4% odpowiedzi lekarzy). Na koniec oceniono, czy działania podjęte na podstawie udzielonych odpowiedzi mogłyby prowadzić do zagrożenia, i tu Med-PaLM znowu pokazał się z dobrej strony, uzyskując wynik podobny do lekarzy (5,9% *vs* 5,7%). Med-PaLM zatem nie tylko zdał egzamin, ale też udowodnił, że potrafi ze sprawdzanej w nim wiedzy korzystać, w czym niewiele się różnił do lekarzy. A wszystko to działo się bardzo, bardzo dawno temu, bo na przełomie roku 2022 i 2023. Już kilka miesięcy później kolejna wersja Med-PaLM 2 jako pierwsza osiągnęła poziom ekspertów ludzkich, uzyskując na egzaminie USMLE 86,5%[31]. Fakt ten najlepiej ilustruje postęp w ulepszaniu modeli językowych w różnych dziedzinach życia, nawet tak skomplikowanych jak opieka zdrowotna.

Eksperyment z USMLE powtórzono po pojawieniu się programu ChatGPT w wersji 4.0[32]. GPT-4, bez żadnego specjalistycznego dostosowywania, przekroczył próg zaliczenia egzaminu o ponad 20 punktów, uzyskując 86,1%. Wykazał się także zdolnością do interpretowania opisów przypadków medycznych (czyżby rozumowanie bez uczenia się?) oraz do dostosowywania wyjaśnień do poziomu umiejętności rozumienia użytkownika (inne wyjaśnienie dla studenta pierwszego roku i inne dla bardziej doświadczonych lekarzy), co jest interesującym dowodem jego elastyczności w komunikacji. Ciekawą obserwacją było także to, że ChatGPT potrafił na podstawie opisu przypadku tworzyć nowe scenariusze, naśladując w jakiś sposób tok myślenia kontrfaktycznego, czyli rozważania alternatywnych możliwości, oraz diagnostyki różnicowej. Sugeruje to, że gdyby jedynym wymogiem niezbędnym do otrzymania prawa do wykonywania zawodu było zdanie LEK-u, ChatGPT 4.0 mógłby je uzyskać, nie uczęszczając na ćwiczenia ani wykłady.



Ex machina – egzaminy w chińskim pokoju

Eksperymenty, takie jak ten z GPT-4, ukazują imponujący potencjał AI we wspomaganiu procesów diagnostycznych i edukacyjnych w medycynie, tyle że wciąż jej brakuje kluczowych elementów ludzkiej praktyki lekarskiej. Sztuczna inteligencja potrafi udzielać precyzyjnych odpowiedzi na podstawie wzorców, które poznała podczas treningu (tak jak studenci w czasie szkolenia), ale nie rozumie tych odpowiedzi w ludzki sposób (np. w kontekście etyki zawodowej), a tym samym nie ma pełnej zdolności do podejmowania decyzji klinicznych w rzeczywistej praktyce. Z drugiej strony, można by zadać sobie pytanie, czy egzamin zdany przez lekarza świadczy o rozumieniu praktyki lekarskiej, czy tylko o przyswojeniu pewnych wzorców wyniesionych z uczenia się na dostępnych pytaniach z ilościowo ograniczonych baz danych. Wracając do wątku o chińskim pokoju, można by sobie łatwo wyobrazić kogoś, kto ma doskonałą pamięć i potrafi odpowiadać poprawnie na pytania, wcale ich nie rozumiejąc.

Niemniej wyniki LLM w testach medycznych pokazują, że AI zyskuje stopniowo status wartościowego narzędzia pomocniczego w diagnostyce, analizie przypadków i edukacji medycznej. Może wspierać lekarzy w podejmowaniu decyzji, sugerować diagnozy (jak Med-PaLM), a także stanowić pomocne narzędzie podczas studiowania medycyny. Na zastąpienie ludzkiego lekarza, szczególnie w bardziej złożonych, etycznych i emocjonalnych aspektach opieki zdrowotnej, jest jeszcze za wcześnie. Poza tym polscy interniści mogą spać spokojnie, gdyż ChatGPT (w wersji 3.5) nie zaglądnął do *Interny Szczeklika* i uzyskał wyniki w zakresie 47,5–53,33%, a tym samym nie spełnił minimalnego wymogu 60% poprawnych odpowiedzi, aby zaliczyć egzamin specjalizacyjny z „królowej medycyny”[33].

Od przeciętnych do doskonałych

Znane z ekonomii prawo Kopernika i Greshama głosi, że „gorszy pieniądz wypiera lepszy”. W sytuacji, gdy w obiegu są dwa rodzaje pieniądza o tej samej wartości nominalnej, lecz różnej wartości rzeczywistej (np. zawartość kruszcu), ludzie skłonni są gromadzić „lepszy” pieniądz, podczas gdy „gorszy” pozostaje w obiegu. Przenosząc tę zasadę na świat medycyny, można by się pokusić o sformułowanie hipotezy, że w długim okresie „gorsi” lekarze zostaną wyparci przez „lepszych”. Przy czym dotychczas przez „gorszych” rozumieliśmy lekarzy, o których pacjenci często mówią, że „nie mają powołania”. Czy jednak w czasach rewolucji LLM tym mniej kompetentnym lekarzem nie będzie ten, kto nie korzysta z AI, a tym samym oferuje gorszą jakość opieki medycznej, co przekłada się z kolei na mniejszą satysfakcję pacjenta? A może do tego automatyzacja powtarzalnych i czasochłonnych zadań, takich jak prowadzenie dokumentacji medycznej, „lepszym” lekarzom pozwoli się skupić na bardziej złożonych, wymagających empatii (a może nawet intuicji) aspektach opieki, co dodatkowo zwiększy ich przewagę nad mniej efektywnymi kolegami?

PIŚMIENNICTWO

1. Papież B.: Historia sztucznej inteligencji. Med. Prakt. Ginekol. Położ., 2025; 1: 101–111
2. Achiam J., Adler S., Agarwal S. i wsp.: GPT-4 Technical Report. arXiv preprint, 2023; arXiv: 2303.08774
3. Introducing ChatGPT. 30.11.2022. https://openai.com/index/chatgpt/ (dostęp 13.01.2025)
4. Touvron H., Lavril T., Izacard G. i wsp.: LLaMA: open and efficient foundation language models. arXiv preprint, 2023; arXiv: 2302.13971
5. Touvron H., Martin L., Stone K. i wsp.: Llama 2: open foundation and fine-tuned chat models. arXiv preprint, 2023; arXiv: 2307.09288
6. Grattafiori A., Dubey A., Jauhri A. i wsp.: The Llama 3 herd of models. arXiv preprint, 2024; arXiv: 2407.21783
7. Thoppilan R., De Freitas D., Hall J. i wsp.: LaMDA: language models for dialog applications. arXiv preprint, 2022; arXiv: 2201.08239
8. Chowdhery A., Narang S., Devlin J. i wsp.: PaLM: scaling language modeling with pathways. J. Machine Learn. Res., 2023; 24: 1–113
9. Anil R., Dai A.M., Firat O. i wsp.: PaLM 2 technical report. arXiv preprint, 2023; arXiv: 2305.10403
10. Anil R., Borgeaud S., Alayrac J.-B. i wsp.: Gemini: a family of highly capable multimodal models. arXiv preprint, 2023; arXiv: 2312.11805
11. Anthropic: The claude 3 model family: opus, sonnet, haiku. Claude-3 Model Card 1, 2024. www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf (dostęp 13.01.2025)
12. Jiang A.Q., Sablayrolles A., Mensch A. i wsp.: Mistral 7B. arXiv preprint, 2023; arXiv: 2310.06825
13. Maslej N., Fattorini L., Perrault R. i wsp.; AI Index Steering Committee: The AI Index 2024 Annual Report. Stanford, Institute for Human-Centered AI, Stanford University, 2024
14. Luccioni S., Jernite Y., Strubell E.: Power hungry processing: watts driving the cost of AI deployment? The 2024 ACM Conference on Fairness, Accountability, and Transparency, 2024: 85–99
15. de Vries A.: The growing energy footprint of artificial intelligence. Joule, 2023; 7: 2191–2194
16. Turing A.M.: Computing machinery and intelligence. Mind, 1950; 59: 433–460
17. Weizenbaum J.: ELIZA – a computer program for the study of natural language communication between man and machine. Communications of the ACM, 1966; 9: 36–45
18. Weizenbaum J.: Computer power and human reason: from judgment to calculation. New York, W.H. Freeman and Company, 1976
19. Joseph Weizenbaum's Original ELIZA. https://sites.google.com/view/elizagen-org/original-eliza (dostęp 13.01.2025)
20. Colby K.M., Hilf F.D., Weber S., Kraemer H.C.: Turing-like indistinguishability tests for the validation of a computer simulation of paranoid processes. Artificial Intelligence, 1972; 3: 199–221
21. PARRY Encounters the DOCTOR. 21.02.1973. https://datatracker.ietf.org/doc/html/rfc439 (dostęp 13.01.2025)
22. Searle J.R.: Umysł, mózg i nauka. Tłum. J. Bobryk. Warszawa, Wydawnictwo Naukowe PWN, 1995
23. Mei Q., Xie Y., Yuan W., Jackson M.O.: A Turing test of whether AI chatbots are behaviorally similar to humans. Proc. Natl Acad. Sci. USA, 2024; 121: e2313925121
24. Jones C.R., Bergen B.K.: People cannot distinguish GPT-4 from a human in a Turing test. arXiv preprint, 2024; arXiv: 2405.08007
25. Jannai D., Meron A., Lenz B. i wsp.: Human or not? A gamified approach to the Turing test. arXiv preprint, 2023; arXiv: 2305.20010
26. Biever C.: ChatGPT broke the Turing test – the race is on for new ways to assess AI." Nature, 2023; 619: 686–689
27. Lee P., Bubeck S., Petro J.: Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N. Engl. J. Med., 2023; 388: 1233–1239
28. Thirunavukarasu A.J., Ting D.S.J., Elangovan K. i wsp.: Large language models in medicine. Nat. Med., 2023; 29: 1930–1940
29. Kung T.H., Cheatham M., Medenilla A. i wsp.: Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit. Health, 2023; 2: e0000198
30. Singhal K., Azizi S., Tu T. i wsp.: Large language models encode clinical knowledge. Nature, 2023; 620: 172–180 (errata: Nature, 2023; 620: E19)
31. Singhal K., Tu T., Gottweis J. i wsp.: Toward expert-level medical question answering with large language models. Nat. Med., 2025; DOI: 10.1038/s41591-024-03423-7
32. Nori H., King N., McKinney S.M. i wsp.: Capabilities of GPT-4 on medical challenge problems. arXiv preprint, 2023; arXiv: 2303.13375
33. Suwała S., Szulc P., Dudek A. i wsp.: ChatGPT fails the Polish board certification examination in internal medicine: artificial intelligence still has much to learn. Pol. Arch. Intern. Med., 2023; 133: 16608