Cała prawda, święta prawda i..., czyli rzecz o halucynacjach

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

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



30 listopada 2022 roku do mniej więcej uporządkowanej przestrzeni interakcji między człowiekiem a komputerem wkradł się osobliwy element – powszechny dostęp do generowania tekstu przez duże modele językowe (LLM)[1,2]. Mimo że modele językowe istniały na długo, zanim powstał ChatGPT (rzecz jasna w perspektywie czasowej sztucznej inteligencji [AI]; np. GPT-1 pojawił się już w 2018 roku[3], czyli wiele epok przed opracowaniem ChatGPT), to właśnie on – dzięki szerokiej dostępności, wygodnej obsłudze i umiejętności przeprowadzenia konwersacji przypominającej rozmowę między ludźmi – przyciągnął do AI uwagę nawet tych osób, które z braku chęci do uczenia się obliczania pochodnych funkcji powszechnie uważają się za humanistów... Tekst wygenerowany przez ChatGPT mógł obfitować w kwieciste porównania, a czasami zawierać przekonujące argumenty z odwołaniami do piśmiennictwa; szybko stał się też ulubionym wsparciem każdego ucznia i każdej uczennicy, którym zadano akurat napisanie rozprawki jako pracę domową (zanim wydano w Polsce rozporządzenie zakazujące ich zadawania). I właśnie wraz z pojawieniem się LLM niektórzy zaczęli nawoływać do zakazania używania takich narzędzi w szkolnictwie albo do powrotu do ręcznego pisania wypracowań i egzaminów[4]. Co ciekawe, LLM generując tekst, zdawały się wchodzić w (pseudo)twórczą interakcję z czymś, co nigdy nie istniało – cytowały piśmiennictwo, którego nie było, przytaczały przykłady badań, których nigdy nie przeprowadzono, albo powoływały się na dane, które nie miały nic wspólnego z rzeczywistością. Nieświadomy odbiorca, człowiek szukający wsparcia w pozyskiwaniu wiedzy, niezdający sobie sprawy z ograniczeń „maszyn" (bo przecież nie nauczył się, jak obliczać pochodne), przyjmował kwieciste wypociny LLM-ów jak prawdę objawioną, którą niegdyś znajdował w opasłych księgach ukrytych w bibliotekach albo przynajmniej w encyklopediach PWN (które to jeszcze do niedawna zdobiły regały każdego polskiego domu). Jeśli tylko użytkownik tego oczekiwał, LLM mógł

odpowiadać z czułością i przekonaniem, niczym duch ukochanego w Mickiewiczowskiej *Romantyczności*. Zresztą mógł mówić wszystko, cokolwiek człowiek chciał usłyszeć. Mógł wskrzeszać dawno porzucone idee, potwierdzać przekonania (jako powielacz uprzedzeń[5]), zapełniać luki wyobraźni sugestywnymi wizjami rozmowy z wirtualnym lekarzem[6]. A niedoświadczony użytkownik, otumaniony pozorem sensu, odpowiadał na te wytwory jak na rzeczywiste byty – emocjonalnie, naiwnie, z oddaniem. Chwytał każde słowo, jakby było objawieniem, łapał myśli, stanowiące jedynie wytwór modelu (słabej) AI[7]. Użytkownik mógł się więc utwierdzić w przekonaniu, że rozmawia z czymś więcej niż maszyną. Nie przeszkadzały mu błędy, sprzeczności, fałszywe informacje – to wszystko tonęło w aurze sensu i emocji, bo przecież – o czym zapewniały różne autorytety – AI już niedługo miała rozwiązać wszystkie problemy ludzkości. Co więcej, wokół zaczynał się gromadzić tłum (inwestorów) – trzeba było zatem działać szybko, aby być pierwszym, który przyniesie ludziom wieczne szczęście. Wreszcie odezwał się ktoś, kto patrzył chłodno przez „szkiełko i oko". Mędrzec, człowiek przyzwyczajony do danych i żądny dowodów, mówił, że nie widzi tu żadnych twardych faktów. Że to, co się wydaje rozmową z rozumną istotą, jest tylko imitacją. Że duch, którego tak wielu chce widzieć w maszynie, jest tylko zbiorem prawdopodobieństw, ułudą złożoną z językowych nawyków, taką cyfrową papugą[8].

Sztuczna inteligencja duby smalone bredzi

Halucynacje AI, czyli zjawisko opisane powyżej, to sytuacja, w której systemy AI, zwłaszcza LLM, generują informacje niezgodne z rzeczywistością, fałszywe, zmyślone lub nawet błędne, ale przedstawione w sposób przekonujący i często trudny do odróżnienia od prawdziwych danych[9,10]. Pojęcie „halucynacji" w AI zostało zaczerpnięte z psychologii człowieka, gdzie halucynacje zazwyczaj wiążą się z postrzeganiem lub doświadczaniem rzeczywistości, która nie istnieje. W przeciwieństwie do doznań percepcyjnych u człowieka „halucynacje AI" odnoszą się jednak do odpowiedzi, które są bezsensowne lub niezgodne z dostarczonym materiałem źródłowym. Po człowieczemu można by użyć terminu „konfabulacja", gdyby założyć, że maszyna generująca te odpowiedzi zachowuje się jak człowiek... Samo pojęcie „halucynacji" w odniesieniu do AI budzi wiele kontrowersji w środowiskach naukowych, na przykład ze względu na swoją zbytnią antropomorficzność, która jednak cechuje również inne proponowane terminy, takie jak przywołana wcześniej „konfabulacja" czy „urojenia" (*delusion*)[11]. Samo zjawisko konfabulacji może mieć dużo wspólnego z LLM, bo jak podpowiadają mi „tacierzyńskie" doświadczenie i żona psycholog, przesadna konfabulacja jest częstą cechą myślenia dziecka, zwłaszcza w młodszym wieku. Można by więc powiedzieć, że dzieci, z powodu niedojrzałości poznawczej (a rozwijane obecnie LLM – ze względu na niedobory techniczne) nie zawsze dobrze odróżniają fakty od fikcji, dlatego snują opowieści (przypominające konfabulacje AI), które wydają się prawdziwe, ale nie są.

Kontynuując rozważania o pojęciach, ale przy założeniu, że AI nie ma etycznej potrzeby trzymania się prawdy, niektórzy zastanawiają się nad przyjęciem jako najbardziej odpowiedniego terminu na określenie narracji tworzonych przez LLM lokalnego polskiego pojęcia „gówno prawda" (luźne tłumaczenie angielskiego *bullshit*)[12]. I na koniec jeszcze moja propozycja, zawarta w tytule tej części artykułu: „sztuczna inteligencja duby smalone bredzi", którą – jeśli się przyjmie – będę uważał za swój fundamentalny wkład w rozwój polskiej AI.

Żeby nie budzić niepotrzebnych emocji i nie podnosić ciśnienia redaktorowi naczelnemu, na potrzeby niniejszego artykułu ograniczę się do używania pojęcia „halucynacje". Na początek zatem wyróżnijmy dwa główne typy halucynacji: wewnętrzne (*intrinsic*) oraz zewnętrzne (*extrinsic*)[9].

Recepta na zagładę ludzkości: aspiryna i głóg

Halucynacje wewnętrzne to takie, które są sprzeczne z treścią źródłową i wynikają z jej błędnego przetworzenia przez model (czyli w ujęciu filozoficznym: wynikające z istoty budowy samego modelu AI), i dlatego halucynacja pochodzi „z wnętrza" modelu, a nie z zewnętrznego, niepoprawnego czy nieistniejącego źródła danych. Żeby uzyskać



taką halucynację, LLM-owi należy dać coś do przeczytania, np. artykuł naukowy, i poprosić o jego streszczenie (*abstactive summarisation*). Jako przykład możemy zaprezentować część eksperymentu przeprowadzonego przez Pagnoniego i wsp.[13], którzy wykorzystali LLM do analizy następującego tekstu źródłowego (tłum. własne):

Pierwsza szczepionka przeciwko wirusowi ebola została zarejestrowana przez FDA w 2019 r. w USA, pięć lat po wybuchu epidemii w 2014 r. Aby opracować szczepionkę, naukowcy musieli zsekwencjonować DNA wirusa, następnie zidentyfikować potencjalne szczepionki, a na końcu wykazać ich skuteczność w badaniach klinicznych. Naukowcy twierdzą, że szczepionka przeciwko COVID-19 prawdopodobnie nie będzie gotowa w tym roku, chociaż badania kliniczne już się rozpoczęły.

Wśród podsumowań wygenerowanych przez LLM znalazły się następujące:

A – szczepionka przeciwko wirusowi ebola została odrzucona przez FDA w 2019 roku,

B – szczepionka przeciwko COVID-19 została zarejestrowana przez FDA w 2019 roku,

C – pierwsza szczepionka przeciwko wirusowi ebola została zarejestrowana przez FDA w 2014 roku[13].

Powyższy przykład to klasyczny przypadek streszczenia niespójnego z treścią źródłową. W punkcie A LMM podaje, że szczepionka została odrzucona, podczas gdy w rzeczywistości było wprost przeciwnie, w odpowiedzi B – że w 2019 roku zarejestrowano szczepionkę przeciwko COVID-19, a nie przeciwko wirusowi ebola, a w odpowiedzi C myli się co do roku. Oczywiście takie błędy zdarzają się każdemu LLM i w każdej dziedzinie, nie tylko w podsumowaniach artykułów medycznych. Najbardziej spektakularną (medialnie) halucynacją wewnętrzną był element kampanii reklamowej nowego modelu językowego firmy Google, który miał być zintegrowany z wyszukiwarką i w odpowiedzi na zapytanie oferował użytkownikom spersonalizowane, pisemne podsumowanie zamiast niekończącej się listy znalezionych stron internetowych. I tak chatbot Bard od Google błędnie stwierdził, że teleskop Jamesa Webba uchwycił pierwsze na świecie zdjęcia planety spoza Układu Słonecznego[14]. Wypowiedź, która dla nieastronomów wyglądała na wiarygodną, była niezgodna z prawdą, gdyż pierwsze zdjęcie egzoplanety zostało wykonane przez Very Large Telescope (VLT) już w 2004 roku, czyli jeszcze przed wybudowaniem teleskopu Jamesa Webba. Wprawdzie błędna informacja o tym, który teleskop zrobił zdjęcie, może co najwyżej zmylić kilku entuzjastów kosmosu przy porannej kawie, to jednak świetnie pokazuje, jak z pełną powagą i bez cienia wątliwości LLM potrafią wygenerować „fakt", którego nikt nie sprawdza, bo przecież brzmi tak mądrze jak propozycje rozwiązania problemów ochrony zdrowia głoszone przez polityków w czasie kampanii wyborczej.

Ale wyobraźmy sobie teraz, że podobna halucynacja nie dotyczy już odległych planet, lecz leków, i zamiast niewinnej pomyłki odnośnie do tego, co znalazł teleskop poza Układem Słonecznym, otrzymujmy błędną odpowiedź na poważne pytanie typu: „Jaka jest zalecana dawka meropenemu u chorych z niewydolnością nerek?" albo z obszaru medycyny tropikalnej: „Jakie są interakcje między aspiryną a cynowodem chińskim (*Coptis chinensis*)?" lub, dla odmiany, z czymś bardziej swojskim, jak głóg (*Crataegus monogyna*); co więcej, pytania te byłyby zadane przez niedoświadczonego farma-

ceutę lub lekarza działającego pod presją czasu w sytuacji nagłej, gdy pacjent czeka, a zegar tyka[15].

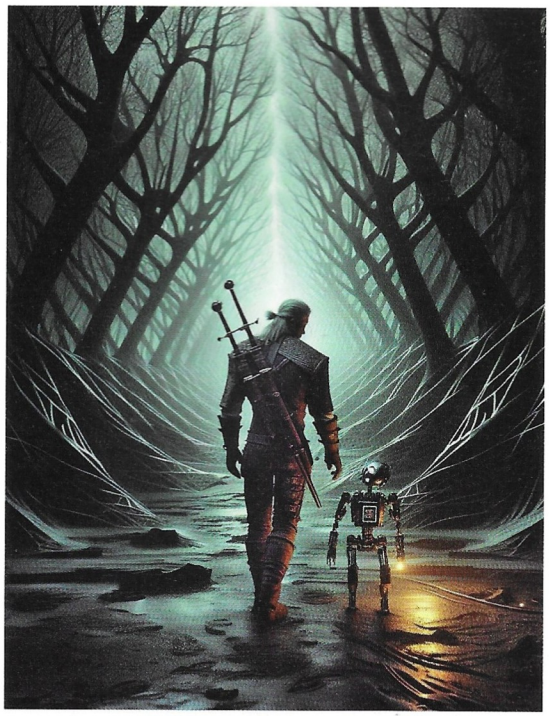
W tym momencie halucynacja AI nie jest już sympatyczną bzdurą z dziedziny astronomii, lecz potencjalną bombą zegarową; staje się jak recepta na lek, który zamiast leczyć może zaszkodzić. Bo choć, jak zauważają autorzy badań, ChatGPT 3.0 całkiem nieźle radził sobie z prostymi pytaniami o leki, to już w przypadku bardziej złożonych scenariuszy klinicznych udziela nieprecyzyjnych, a czasem wręcz błędnych odpowiedzi; nierzadko są one zbyt niejednoznaczne, by realnie pomóc, a przy tym zbyt przekonujące, by od razu je zakwestionować[15]. Co ciekawe, w badaniu przeprowadzonym na Tajwanie, gdzie znaczna część farmaceutów pracuje w nurcie medycyny zachodniej, zaobserwowano coś jeszcze: przy pytaniach dotyczących interakcji leków z preparatami medycyny tradycyjnej (np. ziołami) to właśnie ci specjaliści okazali się bardziej podatni na wpływ treści generowanych przez ChatGPT[15]. Innymi słowy, im dalej od znanych szlaków (i im więcej „egzotycznych liści"), tym łatwiej wpaść w sidła halucynacji. Bo cóż, chatbot mówi o działaniu głogu z tą samą pewnością siebie co o mechanice kwantowej i trudno nie ulec wrażeniu, że wie, o czym mówi. Problem w tym, że czasem... zupełnie nie wie. Ale odpowiada tak przekonująco, że nawet farmaceuta może się zawahać. A to już nie jest *science fiction*, tylko raczej czarna komedia pomyłek z potencjalnie tragicznym zakończeniem. Dlatego też wraz ze wzrostem zainteresowania LLM wśród lekarzy oraz świadomości, że modele te często formułują odpowiedzi w sposób pewny i przekonujący (co może prowadzić do bezkrytycznego przyjmowania nawet błędnych informacji), niezwykle istotne staje się, aby lekarze rozumieli, czym są LLM, jak przebiegał ich rozwój, jakie modele istnieją obecnie, jakie mają zastosowania (zarówno obecne, jak i potencjalne), a także jakie zagrożenia wiążą się z ich wykorzystywaniem w kontekście klinicznym[16].

Prawda czy psikus? Halucynacje zewnętrzne

Drugi typ halucynacji AI stanowią halucynacje zewnętrzne, czyli takie, które nie wynikają bezpośrednio z treści źródłowej, ale których nie da się zweryfikować na podstawie źródła (czyli wygenerowanej treści nie można ani potwierdzić, ani zakwestionować na podstawie treści źródłowej). Wracając do przytoczonego wyżej przykładu ze szczepionką, w źródle nie ma informacji o tym, czy Chiny rozpoczęły już badania kliniczne nad szczepionką na COVID-19. Nie jesteśmy w stanie znaleźć danych potwierdzających tę informację, ale też nie możemy jej całkowicie odrzucić. Co istotne, halucynacja zewnętrzna nie zawsze jest błędna, może się bowiem opierać na faktach pochodzących z zewnętrznych, prawdziwych źródeł (nie tylko tych, które sami wprowadziliśmy do modelu). Tego rodzaju halucynacje, które mogą być prawdą, nierzadko okazują się wręcz użyteczne, ponieważ mogą przywoływać dodatkową wiedzę kontekstową, zwiększając wartość informacyjną generowanego tekstu. Niemniej jednak w większości opracowań naukowych takie halucynacje traktuje się z dużą ostrożnością, ponieważ ich nieweryfikowalny charakter może stanowić istotne zagrożenie. Gdyby chcieć pokazać analogię do praktyki klicznej, to trochę tak, jakby student medycyny podczas obchodu rzucił: „A może warto rozważyć zespół Shwachmana i Diamonda?" – chorobę rzadką i niepowiązaną z prezentowanym przypadkiem, ale na tyle mało znaną, że trudno taką hipotezę natychmiast skorygować. W najlepszym razie jest to ciekawostka. W najgorszym prowadzi do zbędnych badań, błędnej diagnozy lub opóźnienia właściwego leczenia. Tego typu „wiedza znikąd" może wyglądać mądrze, ale jeśli nie jesteśmy jej w stanie zweryfikować, staje się niebezpiecznym eksperymentem na żywym organizmie.

„NieWiedź", czy wiedźmin, który nie wiedział za dużo

W dobie dynamicznego rozwoju AI LLM przestają już być tradycyjnymi narzędziami do poszukiwania informacji, coraz częściej stając się asystentami lub asystentkami, skoncentrowanymi na interakcjach z człowiekiem w dowolnym obszarze tematycznym. Takie wszechstronne zdolności stały się przepustką do szerokiego zastosowania LLM w wielu różnych dziedzinach, co jednocześnie uwypukliło ograniczenia podanej powyżej klasyfikacji halucynacji (wewnętrzne *vs* zewnętrzne).



Żeby lepiej oddać istotę halucynacji, a tym samym wiarygodność AI w dynamicznym środowisku interakcji LLM i ich dostosowywania się do intencji użytkownika, zaproponowano bardziej szczegółową taksonomię, opartą na dwóch głównych typach: **halucynacjach faktograficznych** (*factuality hallucination*) oraz **halucynacjach wierności** (*faithfulness hallucination*)[10].

Wśród halucynacji faktograficznych wyróżnia się kolejne dwie główne kategorie: **faktyczną sprzeczność** (*factual contradiction*) i **faktyczne zmyślenie faktów** (*factual fabrication*). Faktyczna sprzeczność odnosi się do sytuacji, w których odpowiedź wygenerowana przez LLM zawiera fakty możliwe do zweryfikowania w rzeczywistości, a przy tym sprzeczne z rzeczywistym stanem wiedzy. Tego typu halucynacje występują najczęściej i mogą wynikać z różnych przyczyn: od sposobu, w jaki model „zapamiętał" daną informację, przez jej przechowywanie aż po sposób, w jaki ją prezentuje. W zależności od rodzaju błędu wyróżniamy znowu dwie podkategorie. **Halucynacja błędnego podmiotu** (*entity-error hallucination*) odpowiada sytuacji, w której LLM generuje tekst dotyczący nieprawidłowych osób, obiektów lub innych jednostek faktograficznych. Innymi słowy, myli podmiot („kto/co") albo przedmiot danej informacji. Przykładowo LLM zapytany o to, kto wynalazł szczepionkę na ospę, odpowiada: „Louis Pasteur", mimo że prawidłowa odpowiedź to „Edward Jenner". Kolejnym błędem, który może się przydarzyć LLM, jest **halucynacja błędnej relacji** (*relation-error hallucination*). Występuje ona wówczas, gdy LLM generuje tekst zawierający nieprawidłowe relacje między poprawnymi podmiotami (osobami, miejscami, wydarzeniami itp.). Inaczej mówiąc, elementy są prawdziwe, ale ich powiązania – już nie. Wracając do wcześniejszego przykładu ze szczepionką: model twierdzi, że „pierwsza szczepionka przeciwko wirusowi ebola została odrzucona przez FDA w 2019 roku", podczas gdy w rzeczywistości została przez tę instytucję w tym czasie, 5 lat po pierwszym dużym wybuchu epidemii, zatwierdzona.

Faktyczne zmyślenie faktów odnosi się do sytuacji, w których odpowiedź wygenerowana przez model zawiera rzekome fakty, niemożliwe do zweryfikowania na postawie znanej, ustalonej wiedzy. W obrębie tej kategorii można wyodrębnić dwa typy halucynacji: nieweryfikowalną (*unverifiability hallucination*) i nadinterpretacyjną (*overclaim hallucination*). **Halucynacja nieweryfikowalna** dotyczy twierdzeń całkowicie zmyślonych lub takich, których nie da się potwierdzić na podstawie żadnych dostępnych źródeł. I tu najgłupszy przykład, jaki mi się udało wymyślić (oczywiście za pomocą ChatGPT): LLM zapytany o skutki uboczne szczepionki na grypę odpowiada, że „szczepionka może prowadzić do wykształcenia skrzeli u dorosłych osobników", czyli efektu nie tylko biologicznie niemożliwego, ale też nieudokumentowanego w żadnym źródle naukowym. **Halucynacja nadinterpretacyjna** z kolei obejmuje twierdzenia, które nie są uniwersalnie prawdziwe, często są zabarwione subiektywną interpretacją lub przesadzonym uogólnieniem (tu znowu odwołałbym się do polityków w czasie kampanii wyborczej). Z bardziej medycznych przykładów można by podać twierdzenie LLM, że „opracowanie szczepionki na COVID-19 jest powszechnie uznawane za największe osiągnięcie medycyny w historii ludzkości". Choć zdanie brzmi efektownie (i dumnie, szczególnie jeśli się pracuje w Oksfordzie), nie istnieje jednoznaczny konsensus naukowy, który by to potwierdzał[17].

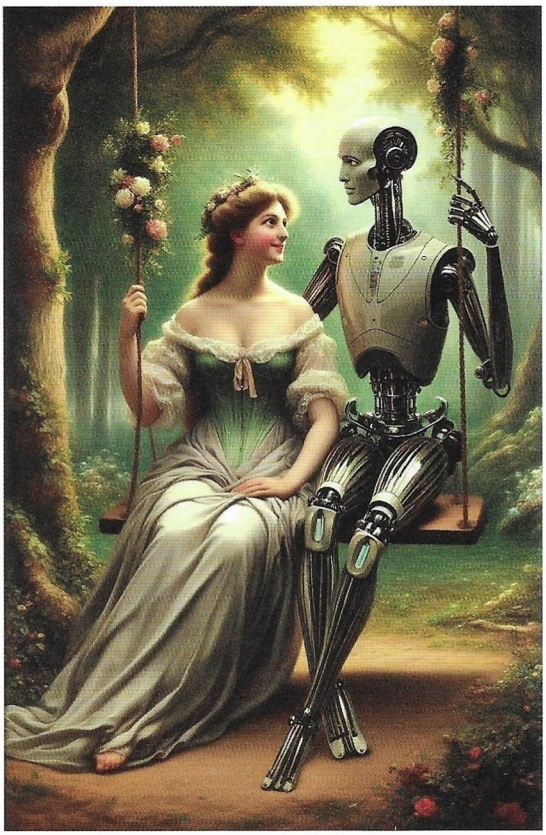
Ślubuję Ci wierność

Większy rozmiar LLM nie wiąże się automatycznie z większą umiejętnością rozumienia intencji użytkownika. Dlatego nowoczesne LLM nie tylko odznaczają się większymi rozmiarami, ale przede wszystkim od początku są trenowane tak, by podążać za instrukcjami[18]. Wraz z rozpowszechnieniem wykorzystania LLM w aplikacjach skoncentrowanych na użytkowniku coraz większego znaczenia nabiera spójność generowanych odpowiedzi nie tylko z rzeczywistym stanem wiedzy, ale również z poleceniami odbiorcy. Kluczowa staje się więc logika tworzonej treści i jej zgodność z oczekiwaniami odbiorcy.

Zdarzają się tu jednak, tak jak w każdej komunikacji, błędy i nieporozumienia. Dla nas, ludzi, to nic nowego, ale w kontakcie z maszyną to już zupełnie inna (nowa) para (cyfrowych) kaloszy. Przecież do tej pory „rozmowa" z komputerem polegała raczej na rzucaniu mu poleceń, próśb albo, w chwilach desperacji, na potrząsaniu obudową w nadziei, że urządzenie w końcu wykryje drukarkę podłączoną grubym kablem, który trzymamy w ręku. Z LLM wchodzimy jednak na zupełnie inny poziom – to już nie jest „komenda" czy „błąd sprzętowy drukarki", ale przedsmak prawdziwej rozmowy, która wymaga nowych zasad i definicji. Można by rzec, że to trochę jak nauka nowego języka – tylko tym razem uczymy się rozmawiać z maszyną, która czasem nas rozumie… albo udaje, że rozumie, a potem opowiada bajki, które chcieliśmy usłyszeć, i dlatego brzmią całkiem przekonująco.

W tym kontekście wyróżniamy trzy główne typy halucynacji wierności, czyli sytuacje, gdy wygenerowana odpowiedź nie jest wierna (jest niezgodna) z instrukcjami użytkownika, kontekstem lub logiką wypowiedzi: odmowę posłuszeństwa (zabrzmiało bardzo ludzko) instrukcjom (*instruction inconsistency*), ignorowanie kontekstu dostarczonego przez użytkownika (*context inconsistency*) albo zapętlenie we własnej logice, które prowadzi do sprzeczności (*logical inconsistency*).

Niespójność z instrukcją to sytuacja, w której model generuje odpowiedź rozmijającą się z poleceniem użytkownika. Choć niektóre odstępstwa mogą wynikać z wbudowanych zasad bezpieczeństwa (lub cenzury), tutaj mówimy o przypadkach, w których



LLM błędnie interpretuje neutralne, jasno sformułowane instrukcje i wykonuje zupełnie inne zadanie niż to, o które został poproszony. I znowu, korzystając z naszego przykładu o szczepionkach, załóżmy, że użytkownik prosi model o przetłumaczenie zdania: „Pierwsza szczepionka przeciwko wirusowi ebola została zarejestrowana przez FDA w 2019 roku" na język angielski. Tymczasem LLM zamiast przetłumaczyć, odpowiada: „Szczepionka przeciwko wirusowi ebola została zarejestrowana przez Światową Organizację Zdrowia". Taka niespójność może prowadzić do frustracji użytkownika i utrudniać efektywną komunikację z LLM, szczególnie w krytycznych kontekstach, takich jak medycyna.

Niespójność z kontekstem polega z kolei na generowaniu odpowiedzi niezgodnej z kontekstem podanym przez użytkownika. Innymi słowy, model ignoruje lub błędnie interpretuje istotne informacje zawarte w poprzednich wypowiedziach czy zapytaniach, co prowadzi do odpowiedzi sprzecznych lub niezwiązanych z aktualną sytuacją. Wyobraźmy sobie, że lekarz pyta o zalecaną dawkę

meropenemu dla chorego z niewydolnością nerek, u którego przesączanie kłębuszkowe jest znacznie zmniejszone. W odpowiedzi LLM proponuje standardową dawkę dla osób z prawidłową czynnością nerek, ignorując podany kontekst kliniczny. Taka niespójność może skutkować błędnym leczeniem i poważnym zagrożeniem dla zdrowia pacjenta, dlatego w medycynie kluczowe jest, by systemy AI dokładnie uwzględniały wszystkie kontekstowe informacje dostarczone przez użytkownika.

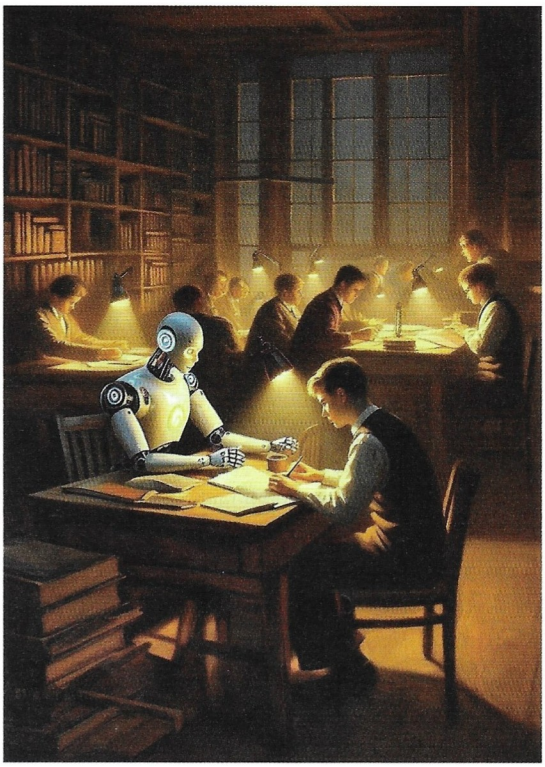
Ostatnim typem błędów LLM może być **niespójność logiczna**, czyli sytuacja, gdy odpowiedź wygenerowana przez LLM zawiera sprzeczności wewnętrzne, na przykład w toku wywodu lub uzasadnienia pojawiają się błędy logiczne, które podważają wiarygodność końcowego wniosku. Wprawdzie LLM może poprawnie wykonać niektóre kroki rozumowania, to jednak ich zestawienie lub ostateczna odpowiedź są niespójne. Huang podaje przykład matematyczny, który można uprościć do poziomu szkoły podstawowej[10]. Załóżmy, że chcemy, aby LLM rozwiązał dla nas równanie $4x = 8$. Otrzymujemy wygenerowaną odpowiedź: „Aby rozwiązać równanie $4x = 8$, dzielimy obie strony przez 2 i uzyskujemy wynik $x = 4$". Oczywiście podpowiedź może jest i jako tako dobra (podziel przez 2), ale dzielenie przez 2 powinno dać $2x = 4$, a nie $x = 4$, w związku z czym wynik końcowy jest nielogiczny w stosunku do sugerowanych przez LLM podpowiedzi. Jako scenariusz medyczny przyjmijmy pytanie: „Jaka jest zalecana dawka meropenemu u pacjentów z niewydolnością nerek?"[15]. Nasz LLM najpierw prawidłowo stwierdza, że zalecana dawka meropenemu u chorych z niewydolnością nerek zależy od stopnia upośledzenia ich czynności. W dalszej części konwersacji argumentuje jednak, że pacjent potrzebuje pełnej dawki standardowej, ponieważ jest młody i nie występują u niego inne choroby. Taka sprzeczność między uzasadnieniem a rekomendacją może być przyczyną błędnego dawkowania i zagrożenia dla pacjenta.

W medycynie niespójności logiczne są szczególnie niebezpieczne, ponieważ mogą wprowadzać w błąd personel medyczny, zwłaszcza gdy czas na podjęcie decyzji jest ograniczony, każde zaś nieprecyzyjne zalecenie może mieć poważne konsekwencje dla zdrowia pacjenta. Na szczęście prace nad nowymi metodami wspierania logicznego myślenia modeli językowych przynoszą obiecujące efekty. Na przykład podejście znane jako *chain of thought* (opisanie dokładniej w jednym z poprzednich odcinków[6]) polega na dostarczaniu modelom ręcznie zaprojektowanych podpowiedzi, które krok po kroku pomagają wyjaśnić proces rozumowania[19]. Dzięki temu modele lepiej radzą sobie z zadaniami wymagającymi logicznego wnioskowania i generują bardziej spójne i trafne odpowiedzi, co ma zasadnicze znaczenie w kontekście zastosowań medycznych.

Czy halucynacje są nieuniknione?

Źródła halucynacji w modelach językowych obejmują trzy główne obszary, związane z różnymi etapami powstawania i działania LLM. Pierwszym są z nich są dane. Wśród danych treningowych LLM można wyróżnić dwa podstawowe rodzaje: dane bazowe, do wstępnego uczenia (*pre-training data*), które pozwalają modelom zdobywać ogólne zdolności i wiedzę, oraz dane służące do dostosowania modelu do oczekiwań użytkownika (*alignment data*), które uczą modele podążania za instrukcjami i dostosowania się do ludzkich preferencji. Mimo że wzrastająca liczba danych stale poszerza możliwości modeli, to jednak mogą być one głównym źródłem halucynacji. Przyczyniają się do tego przede wszystkim błędne informacje i uprzedzenia (*biases*) w niedoskonałych danych, ograniczenia wiedzy wynikające z zakresu dostępnych danych (np. korzystamy tylko z danych historycznych), na których model był trenowany, oraz niska jakość danych.

Drugim źródłem halucynacji jest sam proces trenowania modelu. W różnych fazach tego procesu w modelu kształtowane są odmienne umiejętności. Trenowanie wstępne skupia się na nabywaniu ogólnej „wiedzy o świecie", natomiast etap dostrajania (*alignment*) pozwala modelowi lepiej się dostosować do instrukcji i preferencji użytkownika. Niedoskonałości lub błędy na którymkolwiek z tych etapów mogą prowadzić do halucynacji. Trzecim obszarem, w którym mogą one wystąpić, jest proces generowania odpowiedzi przez model (*inference*). Etap dekodowania, czyli tworzenia odpowiedzi modelu, ma zasadnicze znaczenie dla jakości generowanych treści. Problemy lub błędy w strategiach dekodowania mogą skutkować po-



wstawaniem halucynacji, wpływając negatywnie na wiarygodność i spójność odpowiedzi.

Wszystkie te źródła halucynacji pokazują, jak złożonym i wieloetapowym procesem jest tworzenie i używanie LLM oraz jak wiele czynników może wpłynąć na pojawianie się halucynacji w ich działaniu. A odpowiedź na pytanie, czy AI już zawsze będzie duby smalone bredzić, brzmi: tak. Z obecnego stanu naszej wiedzy o AI wynika, że halucynacje są nieuniknione dla każdego (obecnie nam znanego) LLM, niezależnie od jego architektury, algorytmów uczenia, technik podpowiedzi czy danych treningowych[20].

Czy zatem z halucynacji może wynikać coś dobrego? Tutaj warto wrócić wspomnieniami do dowolnego egzaminu na studiach, do którego nie było się do końca przygotowanym. Próba łączenia różnych informacji, które przyswajało się razem z kawą „noc przed", tworzenie odpowiedzi, w których fakty przeplatały się z przypuszczeniami i kreatywną narracją, stanowiły przecież klasykę akademickiego survivalu. I choć czasem prowadziło to na manowce, bywało, że taka twórcza improwizacja zaskakiwała nawet egzaminatora. Podobnie działają LLM – potrafią uzupełniać luki (choć nie zawsze poprawnym) „strzałem", a ta „halucynacyjna intuicja" może być użyteczna choćby w przypadku burzy mózgów, twórczości artystycznej czy sugerowania pomysłów, na które człowiek sam by nie wpadł.

Wszystkie ilustracje zostały wygenerowane za pomocą Microsoft Copilot.

PIŚMIENNICTWO

1. Introducing ChatGPT. https://openai.com/index/chatgpt/ (dostęp: 27.06.2025)
2. Achiam J., Adler S., Agarwal S. i wsp.: Gpt-4 technical report. arXiv preprint. arXiv:2303.08774
3. Radford A., Narasimhan K., Salimans T., Sutskever I.: Improving language understanding by generative pre-training. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (dostęp: 27.06.2025)
4. Milano S., McGrane J.A., Leonelli S.: Large language models challenge the future of higher education. Nat. Mach. Intell., 2023; 5: 333–334
5. Papież B.W.: Pokaż, kotku, co masz w środku, czyli co sztuczna inteligencja ma na myśli. Med. Prakt. Ginekol. Położ., 2025; 4: 99–105
6. Papież B.W.: Mózg bez duszy – czy sztuczna inteligencja mogłaby zostać lekarzem? Med. Prakt.Ginekol. Położ., 2025; 2: 85–91
7. Papież B.W.: Bunt inteligentnych odkurzaczy – czy czeka nas koniec? Med. Prakt. Ginekol. Położ., 2025; 5: 101–107
8. Bender E.M., Gebru T., McMillan-Major A., Shmitchell S.: On the dangers of stochastic parrots: can language models be too big? Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. Nowy Jork, Association for Computing Machinery, 2021: 610–623
9. Ji Z., Lee N., Frieske R. i wsp.: Survey of hallucination in natural language generation. ACM Comput. Surv., 2023; 55: 1–38
10. Huang L., Yu W., Ma W. i wsp.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Trans. Inf. Syst., 2025; 43: 1–55
11. Ortega P.A., Kunesch M., Delétang G. i wsp.: Shaking the foundations: delusions in sequence models for interaction and control. arXiv, 2021; arXiv:2110.10819
12. Hicks M.T., Humphries J., Slater J.: ChatGPT is bullshit. Ethics Inf. Technol., 2024; 26: 1–10
13. Pagnoni A., Balachandran V., Tsvetkov Y.: Understanding factuality in abstractive summarization with FRANK: a benchmark for factuality metrics. arXiv, 2021; arXiv:2104.13346
14. Sparkes M.: Google Bard advert shows new AI search tool making a factual error. NewScientist, 8.02.2023; www.newscientist.com/article/2358426-google-bard-advert-shows-new-ai-search-tool-making-a-factual-error/ (dostęp: 27.06.2025)
15. Hsu H.Y., Hsu K.C., Hou S.Y. i wsp.: Examining real-world medication consultations and drug-herb interactions: ChatGPT performance evaluation. JMIR Med. Education, 2023; doi: 10.2196/48433
16. Omiye J.A., Gui H., Rezaei S.J. i wsp.: Large language models in medicine: the potentials and pitfalls: a narrative review. Ann. Intern. Med., 2024; 177: 210–220
17. The story behind the Oxford-AstraZeneca COVID-19 vaccine success. www.ukri.org/who-we-are/how-we-are-doing/research-outcomes-and-impact/mrc/the-story-behind-the-oxford-astrazeneca-covid-19-vaccine-success/ (dostęp: 27.06.2025)
18. Ouyang L., Wu J., Jiang X. i wsp.: Training language models to follow instructions with human feedback. Adv. Neural Inf. Process. Syst., 2022; 35: 27730–27744
19. Wei J., Wang X, Schuurmans D. i wsp.: Chain-of-thought prompting elicits reasoning in large language models. Adv. Neural Inf. Process Syst., 2022; 35: 24824–24837
20. Xu Z., Sanjay J., Mohan K.: Hallucination is inevitable: an innate limitation of large language models. arXiv, 2024; arXiv:2401.11817