AI Skills, Powered by Values

values

www.valuesai.eu

MODUŁ 04

Niezawodna treść wspierana sztuczną inteligencją

Źródła, weryfikacja oraz kontrola halucynacji

Co-funded by the European Union

Współfinansowane przez Unię Europejską. Wyrażone poglądy i opinie są wyłącznie poglądami i opiniami autora lub autorów i niekoniecznie odzwierciedlają stanowisko Unii Europejskiej ani Fundacji Rozwoju Systemu Edukacji. Ani Unia Europejska, ani instytucja przyznająca grant nie ponoszą za nie odpowiedzialności.

Zawartość

01	Tworzenie rzetelnych treści w erze sztucznej inteligencji (10 min)	3
02	Zagrożenia związane z treściami sztucznej inteligencji (20 min)	8
03	Odpowiedzialne autorstwo (20 min)	15
04	Praca z materiałami źródłowymi (15 min)	23
05	Halucynacje sztucznej inteligencji (10 min)	36
06	Weryfikacja treści (10 minut)	45
07	Analiza danych z zastosowaniem sztucznej inteligencji (10 min)	51

Co-funded by the European Union

Współfinansowane przez Unię Europejską. Wyrażone poglądy i opinie są wyłącznie poglądami i opiniami autora lub autorów i niekoniecznie odzwierciedlają stanowisko Unii Europejskiej ani Fundacji Rozwoju Systemu Edukacji. Ani Unia Europejska, ani instytucja przyznająca grant nie ponoszą za nie odpowiedzialności.

01

Tworzenie rzetelnych treści w erze sztucznej inteligencji

(10 minut)

Niezawodna zawartość AI

Tworzenie wiarygodnych treści w erze sztucznej inteligencji to nie tylko szybkie pisanie, ale przede wszystkim zapewnienie dokładności, przejrzystości i odpowiedzialności. Wykorzystanie sztucznej inteligencji nie obniża automatycznie jakości, lecz wymaga intensywniejszego nadzoru ze strony człowieka, szczególnie w kontekście faktów, danych, zdrowia, polityki czy kwestii społecznych.

AI Skills,
Powered
by Values
values

Niezawodna treść AI – fundamentalne zasady

- **Podawaj źródła.** Odwołuj się do badań, raportów, oficjalnych oświadczeń oraz oryginalnych danych.
- **Zweryfikuj przed publikacją.** Każdy element treści powinien być dokładnie sprawdzony pod względem faktów, danych liczbowych oraz kontekstu.
- **Ujawniaj zastosowanie sztucznej inteligencji.** Przejrzystość wzmacnia zaufanie odbiorców.
- **Pisz zwięźle i precyzyjnie.** Unikaj niejednoznacznych sformułowań, które mogą prowadzić do błędnej interpretacji.
- **Aktualizuj treść.** W przypadku wystąpienia nowych informacji lub wykrycia błędu, skoryguj materiał.
- **Zadbaj o wyraźne określenie autorstwa i odpowiedzialności.** Treść powinna mieć rozpoznawalnego autora lub redaktora.
- **Nie manipuluj odbiorcami.** Sztuczna inteligencja nie powinna być wykorzystywana do celowego wprowadzania ludzi w błąd, nawet jeśli przyczynia się to do zwiększenia zasięgu.

Niezawodna treść AI – co kształtuje zaufanie

Wiarygodne treści zazwyczaj wyjaśniają nie tylko co, ale również dlaczego i na jakiej podstawie, dlatego połączenie danych z komentarzem eksperta oraz praktycznymi przykładami sprawdza się doskonale.

- Daty wydania i aktualizacji,
- Linki do źródeł pierwotnych,
- Ekspert lub autor dysponujący odpowiednimi kwalifikacjami,
- Logiczna struktura tekstu.
- Wyraźne rozróżnienie pomiędzy faktem, opinią a interpretacją.
- Bez sensacji i języka nastawionego na przyciąganie kliknięć.

W praktyce wiarygodne treści w erze sztucznej inteligencji powstają, gdy sztuczna inteligencja wspiera pracę, ale to ludzie wciąż ponoszą odpowiedzialność za prawdziwość i sens przekazu.

h .

Dlaczego „wiarygodna treść” ma znaczenie w kontekście sztucznej inteligencji

Narzędzia AI są w stanie szybko generować treści, jednak każdy rezultat stanowi jedynie szkic: płynny tekst, wiarygodne dane i przekonujące podsumowania, które mogą okazać się błędne.

Dlaczego to ma znaczenie:

- **Po publikacji, niska jakość treści może zaszkodzić Twojej wiarygodności i wprowadzić odbiorców w błąd.**
- **Sztuczna inteligencja jest zdolna do generowania cytatów, statystyk i wypowiedzi, które wydają się autentyczne (halucynacje).**
- **„Niezawodny” oznacza coś więcej niż biegły – oznacza oparty na źródłach, zweryfikowany i odpowiedzialny.**

Zapytanie do ChatGPT o zdefiniowanie pojęcia skutkuje płynną odpowiedzią, jednak bez wskazania źródeł. To samo pytanie w NotebookLM, z załadowanymi akademickimi plikami PDF, generuje definicję z odniesieniami do rzeczywistych autorów.

02

Ryzyko związane z treściami sztucznej inteligencji

(20 minut)

Treści generowane przez sztuczną inteligencję niosą ze sobą trzy zagrożenia:

Halucynacje, brakujące źródła oraz wykrywanie sztucznej inteligencji

ludzie stosują sztuczną inteligencję nerowania treści, często skupiają się e na szybkości. W rzeczywistości generowane przez sztuczną encję wiążą się z trzema głównymi żeniami.

01

Ryzyko halucynacji (nieprawdziwe informacje):

Sztuczna inteligencja jest zdolna do generowania dat, statystyk, nazw i cytatów, które brzmią autentycznie, lecz nigdy nie miały miejsca.

02

Ryzyko źródłowe (niedostatek podstaw):

Ogólne modele, takie jak ChatGPT 4.0, generują spójny tekst bez odniesień do rzeczywistych publikacji. Bibliografia jest niejasna, niekompletna lub sfabrykowana.

03

Ryzyko identyfikacji (odciski palców AI):

Detektory odnoszą się do ogólnego, listowego, formułowego tekstu generowanego przez AI. Wynik bez edycji osiąga 100% punktów AI, natomiast wynik po edytowaniu spada poniżej 10%.

„Wygląda wiarygodnie”, lecz w rzeczywistości takie nie jest (mit kontra rzeczywistość)

Zwróć uwagę na:

- Idealnie sformatowane cytaty z nieprawidłowymi numerami stron
- Prawdziwe nazwiska autorów połączone z fikcyjnymi dziełami.
- Prawdopodobne statystyki bez potwierdzonego źródła
- Płynne definicje, które są niezgodne z aktualną literaturą.

Mit	***„Jeśli coś ma naukowy charakter, jest wiarygodne.”***
Fakt	Sztuczna inteligencja doskonale imituje styl akademicki, jednocześnie wciąż tworząc nieprawdziwe informacje.
Mit	***„Jeśli detektor zidentyfikuje, że tekst został napisany przez człowieka, jest bezpieczny.”***
Fakt	Detektory analizują jedynie styl. Edytowane dane wyjściowe AI mogą osiągnąć 92% wartości oryginalnej, a mimo to wciąż zawierać halucynacje.

Czym jest AI SLOP?

Błędy AI to niskiej jakości, wprowadzająca w błąd lub niedokładna treść tworzona przez sztuczną inteligencję.

Brzmi to pewnie, ale obejmuje:

- Sfabrykowane fakty
- Błędne numery lub daty
- Nieprawidłowe nazwy, cytaty lub stwierdzenia naukowe
- Powtarzalną, ogólną strukturaę
- „Gładki” tekst pozbawiony jakichkolwiek dowodów

Dlaczego to się dzieje?

01

Model przewiduje wzorce, a nie prawdę – generuje najbardziej prawdopodobną odpowiedź na podstawie danych, ale w rzeczywistości nie „wie”, czy jest ona prawdziwa.

02

Może uzupełniać luki fikcyjnymi szczegółami – w przypadku braku informacji model może dodać coś, co brzmi logicznie, ale niekoniecznie odzwierciedla rzeczywistość.

03

Informacje nie zawsze są aktualne – jeśli wiedza modelu jest ograniczona czasowo, może on przedstawiać nieaktualne fakty lub pomijać niedawne wydarzenia.

04

Może wystąpić błędne zrozumienie kontekstu – niejasne pytanie, skrót lub dwuznaczność mogą prowadzić do niewłaściwej odpowiedzi.

Dlaczego to się dzieje?

05

Brzmi pewnie, nawet gdy jest błędne – sformułowanie może wydawać się niezwykle wiarygodne, mimo że treść zawiera błędy.

06

Może łączyć rzeczywiste fakty z halucynacjami – niektóre elementy odpowiedzi mogą być poprawne, podczas gdy inne są wymyślone lub niepotwierdzone.

07

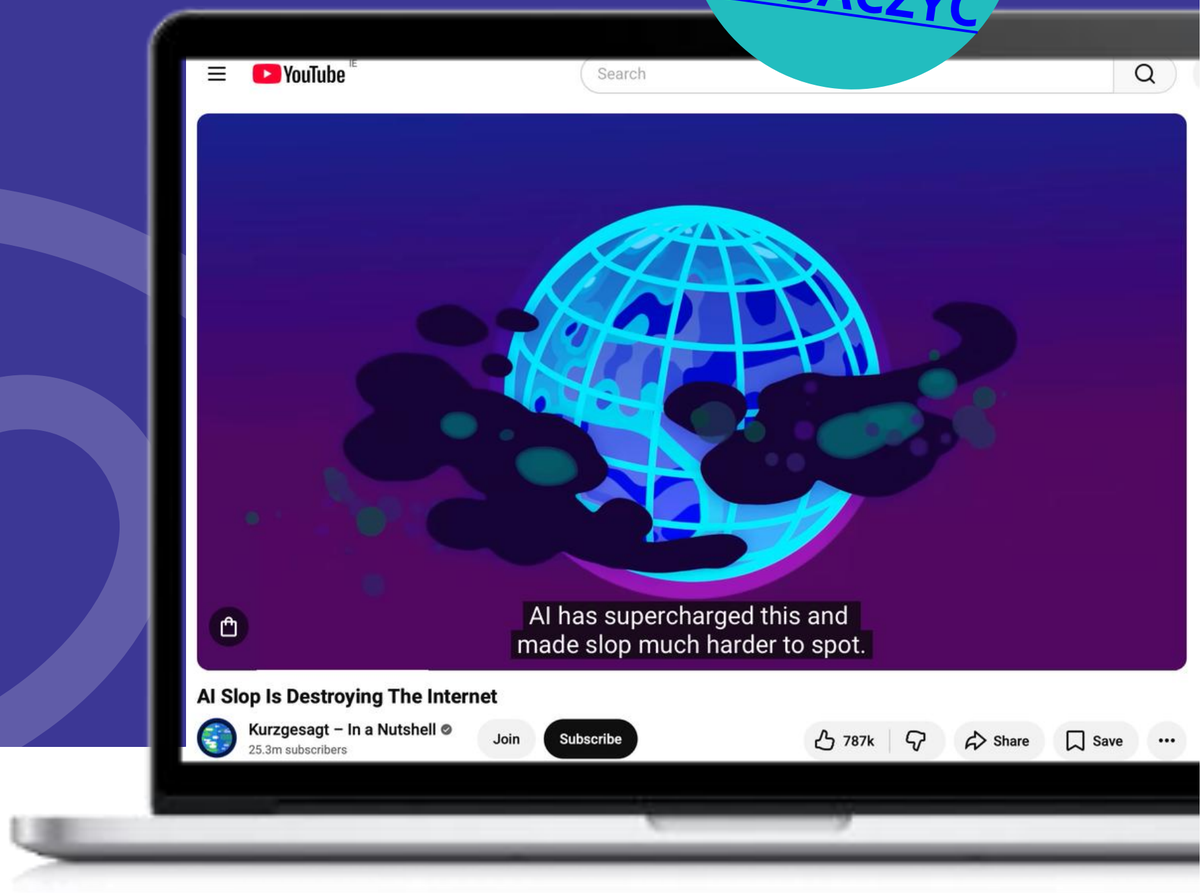
Zależy od jakości danych wejściowych – jeśli zapytanie jest niejasne, sprzeczne lub niekompletne, dane wyjściowe mogą być mniej precyzyjne.

latego weryfikacja staje się fundamentalną umiejętnością cyfrową.

Jak sztuczna inteligencja wpływa negatywnie na Internet

KLIKNIJ, ABY ZOBACZYĆ

Treści generowane przez sztuczną inteligencję zalewają internet – od filmów i muzyki po wiadomości i książki. Co więcej, generatywna sztuczna inteligencja sprawia, że fałszywe informacje wydają się bardziej przekonujące niż kiedykolwiek. Wkraczamy w nową erę przeciążenia informacyjnego, a odróżnianie prawdy od fikcji staje się coraz większym wyzwaniem.



03

Odpowiedzialne autorstwo

(10 minut)

Odpowiedzialny obiektyw.

Za każdym razem, gdy generujesz treści przy użyciu sztucznej inteligencji, zadaj sobie cztery pytania kontrolne:

Dokładność	Źródła	Oryginalność	Odpowiedzialność
Czy zweryfikowałem fakty, daty i cytaty, porównując je z rzetelnymi źródłami?	Czy mogę przedstawić autentyczne publikacje, na których opierają się wszystkie kluczowe twierdzenia?	Czy edytowałem projekt AI w sposób, który uniemożliwia jego wykrycie jako surowego wyniku AI?	Czy mogę przedstawić, w jaki sposób zastosowałem sztuczną inteligencję, co zmieniłem oraz dlaczego?

Rzetelna treść = inteligentne wykorzystanie sztucznej inteligencji z uwzględnieniem odpowiedzialności.

Odpowiedzialny obiektyw.

Bezpieczne korzystanie z AI	Prawa i zgody mają istotne znaczenie.	Sprawiedliwość stanowi integralny element bezpieczeństwa.	Istnieją bardziej bezpieczne przepływy pracy.
To, co wprowadzisz podczas przesyłania, ma kluczowe znaczenie.	Dane osobowe innych osób wymagają zgody oraz ochrony.	Dane stronnicze lub niskiej jakości mogą przynieść szkodę.	Często możesz uzyskać wsparcie sztucznej inteligencji, dzieląc się znacznie mniejszą ilością danych.

Zanim rozpoczniesz korzystanie ze sztucznej inteligencji: każdy rezultat stanowi wersję roboczą.

Wiele osób koncentruje się na szybkości sztucznej inteligencji. Wiarygodne treści zaczynają się wcześniej: od tego, co weryfikujesz, analizujesz i redagujesz.

W praktyce narzędzia sztucznej inteligencji generują:

- **Definicje oraz objaśnienia (często bez odniesień)**
- **Streszczenia prac (czasami fikcyjne lub zniekształcone)**
- **Listy odniesień (często fikcyjne)**
- **Analizy i interpretacje danych (bez metodologii)**

Zanim rozpoczniesz korzystanie ze sztucznej inteligencji:

każdy rezultat stanowi wersję roboczą.

Niezawodne zastosowanie sztucznej inteligencji rozpoczyna się od podstawowego nawyku: traktuj każdy wynik jako wersję roboczą do weryfikacji, a nie jako ostateczną odpowiedź.

AI Skills,
Powered
by Values
values

Wynik AI to zarys, a nie rzeczywistość.

Co uznaje się za treść wymagającą weryfikacji:

- Definicje oraz twierdzenia teoretyczne
- Statystyki, procenty oraz daty
- Nazwy autorów, książek oraz artykułów
- Analizy danych oraz wnioski

CO WOLNO

PAMIĘTAJ: Wszystkie fakty, cytaty i liczby pochodzące z sztucznej inteligencji powinny być zweryfikowane przed użyciem.

CZEGO NIE WOLNO

NIE zakładaj: „To brzmi profesjonalnie” lub „To musi być prawdą, skoro sztuczna inteligencja to stwierdziła”.

Lista
„Nigdy Nie Ufaj”

Jeżeli nie można zidentyfikować źródła, nie publiku

Nie publikuj bez weryfikacji:

- **Cytaty:** wszelkie odniesienia do książki, artykułu lub autora wskazanego przez sztuczną inteligencję.
- **Cytaty:** każdy „bezpośredni cytat” przypisany przez sztuczną inteligencję konkretnej osobie
- **Statystyka**: procenty, liczebność próby, wyniki badań, klasyfikacje
- **Fakty historyczne**: daty, wydarzenia, lokalizacje, nazwy aktów prawnych lub traktatów
- **Definicje i teorie**: stwierdzenia dotyczące tego, czym „jest” koncept lub kto go wprowadził.
- **Dane analizy numerycznej**: wykazy klientów, dane kadrowe, dokumenty wewnętrzne, osoby prywatne

Jak zwiększyć wiarygodność wyników sztucznej inteligencji (przed publikacją)

Przed publikacją wyników sztucznej inteligencji należy przeprowadzić następujące kontrole:

- Sprawdź cytaty: otwórz w Google Scholar, Scopus lub na stronie wydawcy.
- Sprawdź każdą statystykę przynajmniej w jednym niezależnym źródle.
- Uruchom ponownie prompt w zweryfikowanym narzędziu (np. NotebookLM z załadowanymi autentycznymi plikami PDF)
- Edytuj strukturę, zastąp ogólne frazy, wprowadź konkretne przykłady.

żeli nie jesteś w stanie zlokalizować źródła roszczenia → nie publikuj go.

04

Praca z materiałami źródłowymi

(15 minut)

Źródła dostarczone przez sztuczną inteligencję.

01

Korzystaj z rzetelnych źródeł – wybieraj materiały z uznawanych czasopism, instytucji publicznych, baz danych naukowych oraz organizacji eksperckich.

02

Rozróżniamy źródła naukowe, branżowe, medialne oraz opiniotwórcze – każdy z tych typów źródeł pełni odmienną funkcję i zapewnia różny poziom wiarygodności.

03

Weryfikacja dat publikacji – aktualność jest istotna, szczególnie w szybko zmieniających się obszarach, w których sztuczna inteligencja może generować nieaktualne dane.

04

Weryfikacja cytatów – upewnij się, że cytowany autor, tytuł, rok oraz przytoczone stwierdzenie rzeczywiście istnieją i są zgodne z treścią tekstu lub sztucznej inteligencji.

Źródła dostarczone przez sztuczną inteligencję.

05

Rozróżnienie między źródłami pierwotnymi a wtórnymi – źródło pierwotne dostarcza oryginalnego materiału, podczas gdy źródło wtórne analizuje lub interpretuje wcześniejsze prace.

06

Weryfikacja krzyżowa informacji – kluczowe twierdzenia powinny być potwierdzone w co najmniej dwóch lub trzech niezależnych źródłach.

07

Ocena autora, wydawcy i celu – zawsze zastanów się, kto jest autorem tekstu, gdzie został opublikowany oraz czy jego celem jest informowanie, sprzedaż, perswazja czy komentarz.

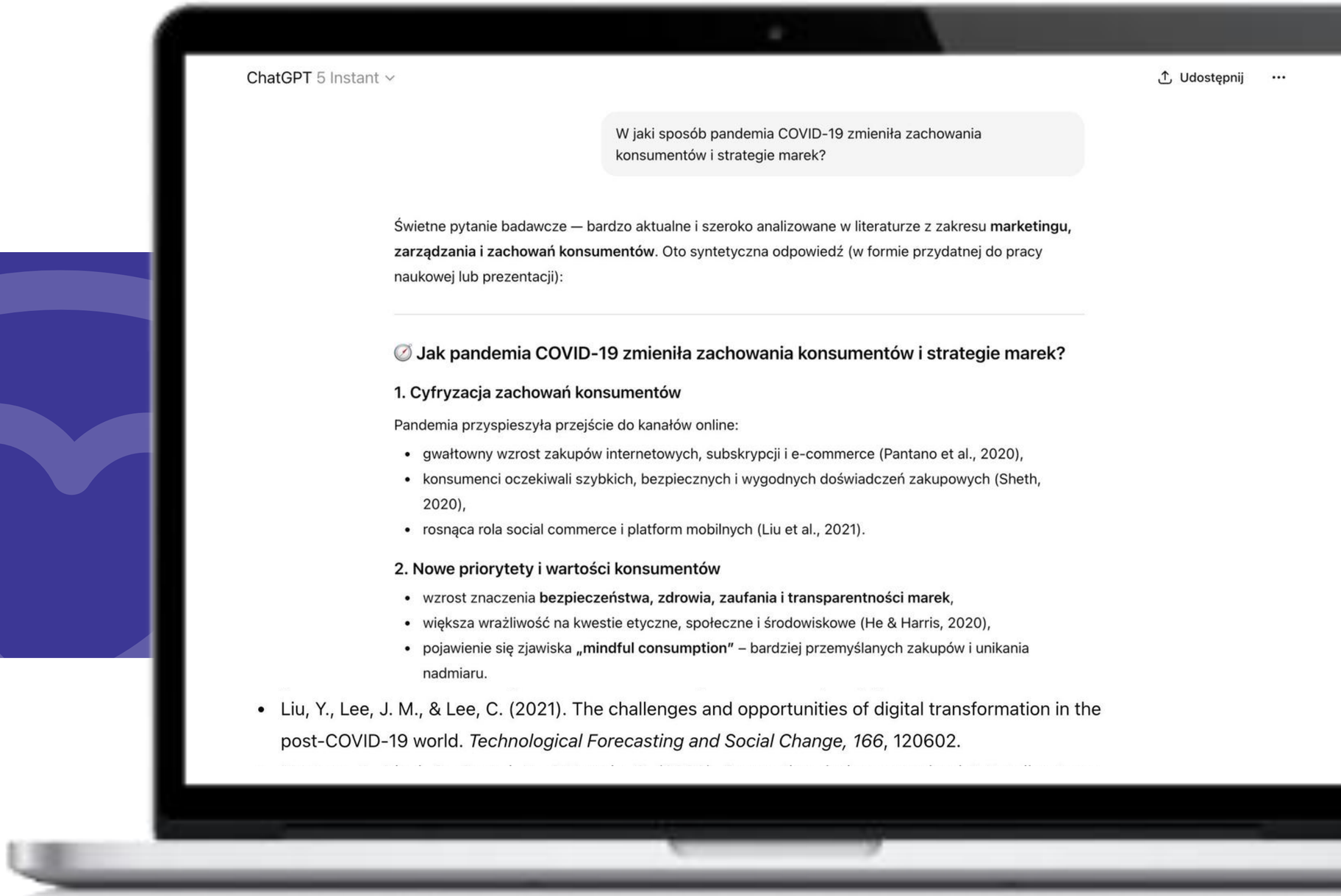
Przykład podania przez sztuczną inteligencję błędnego źródła

Ten slajd ilustruje rozmowę w ChatGPT 5 Instant, w której użytkownik pyta, w jaki sposób pandemia COVID-19 wpłynęła na zachowania konsumentów oraz strategie marek. Sztuczna inteligencja dostarcza zorganizowaną odpowiedź, odwołując się do źródeł akademickich (Pantano i in., 2020; Sheth, 2020; Liu i in., 2021; He i Harris, 2020), analizując digitalizację zachowań konsumenckich oraz nowe priorytety, takie jak bezpieczeństwo, zdrowie, zaufanie i „świadoma konsumpcja”.

Gdy użytkownik zweryfikował jedno cytowanie, odkrył, że artykuł „Wyzwania i możliwości transformacji cyfrowej w świecie po COVID-19” **nie został opublikowany** w czasopiśmie Technological Forecasting and Social Change, jak twierdzi AI, lecz w czasopiśmie IEEE Engineering Management Review, tom 48(3), s. 97–103.

Zobacz kolejny slajd

Przykład podania przez sztuczną inteligencję błędnego źródła

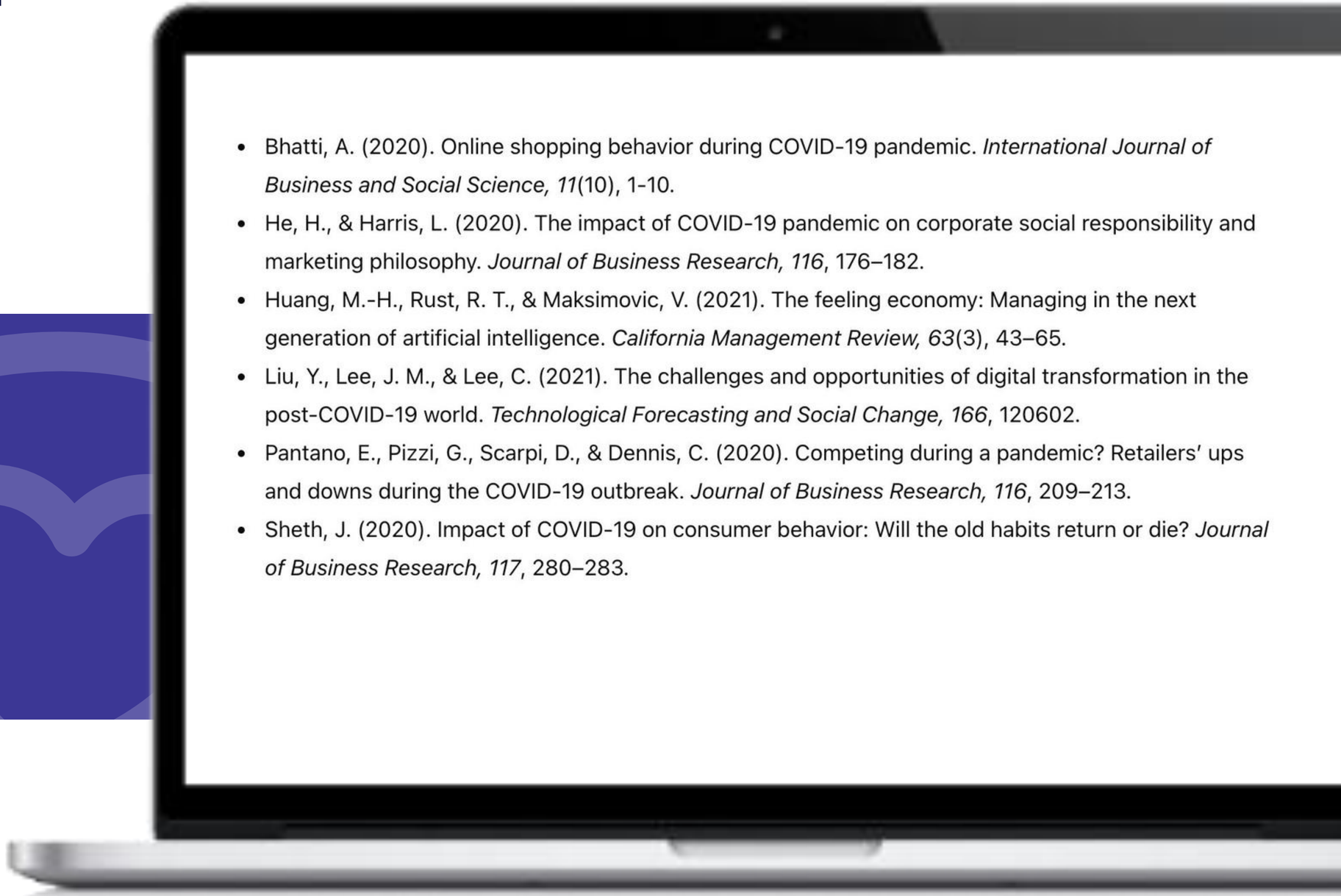


Przykład podania przez sztuczną inteligencję błędnego źródła. Wnioski

- Sztuczna inteligencja błędnie przypisała autentyczny artykuł niewłaściwemu czasopismu, mylnie cytując go jako opublikowany w Technological Forecasting and Social Change zamiast w IEEE Engineering Management Review.
- Wygenerowało to wiarygodny, aczkolwiek fałszywy odnośnik bibliograficzny, zawierający nazwę czasopisma oraz szczegóły publikacji, co doprowadziło do wprowadzenia w błąd w stylu cytowania APA.
- Sztuczna inteligencja nie była w stanie zweryfikować metadanych publikacji (czasopismo, wydawca, DOI), mimo że źródło było łatwe do potwierdzenia.
- Połączono autentyczny tytuł z niewłaściwym kontekstem publikacji, co stanowi typowy przykład halucynacji źródłowej.
- Błąd ten podważa naukową wiarygodność, co może skutkować nieprawidłowymi bibliografiami oraz problemami podczas recenzji eksperckich.

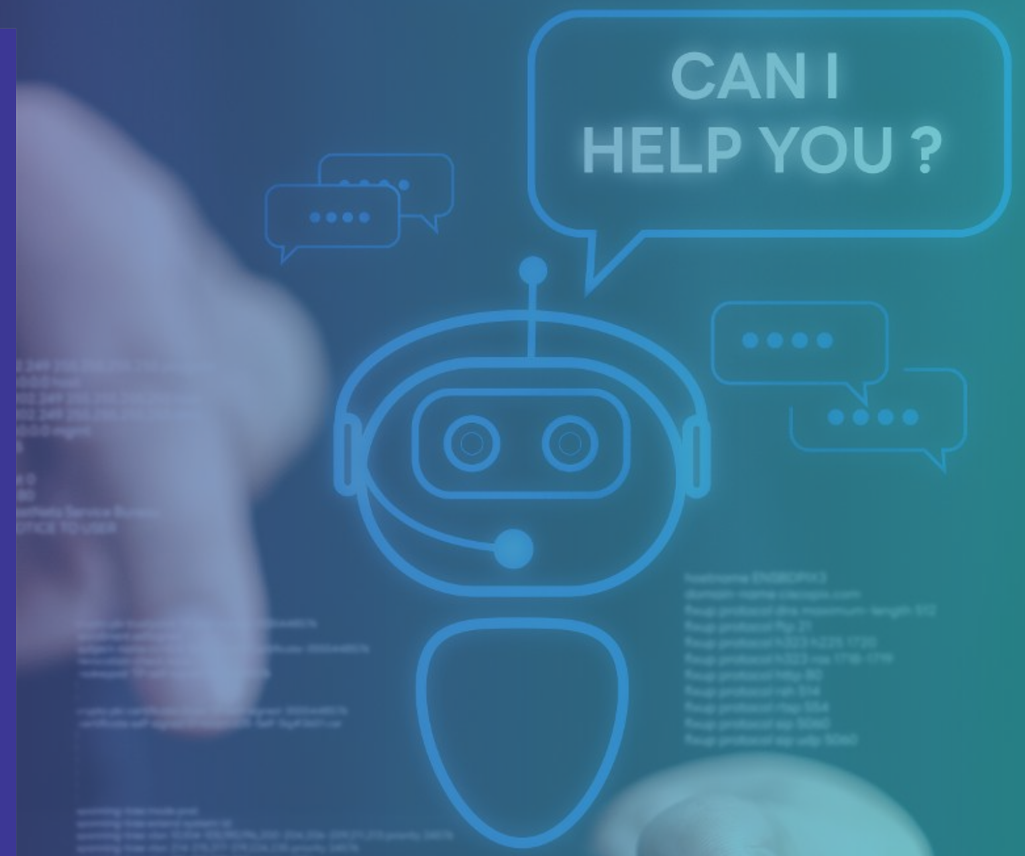
Zobacz kolejny slajd

Przykład podania przez sztuczną inteligencję błędnego źródła. Wnioski



Przykład podania przez sztuczną inteligencję nieprawidłowego źródła. Wnioski

Sztuczna inteligencja może skutecznie podsumowywać literaturę, ale wszystkie generowane przez nią źródła muszą zostać niezależnie zweryfikowane przed wykorzystaniem w pracy naukowej.



AI Skills,
Powered
by Values
values

Ćwiczenia

Poproś dowolny model AI o wygenerowanie jednostronicowego tekstu w formacie A4 na wybrany temat. Zachęć go, aby nie korzystał z internetu, lecz opierał się na publikacjach, takich jak artykuły, książki, raporty i inne podobne źródła.

- Utwórz wykaz źródeł.
- Następnie, zgodnie z najlepszymi praktykami, zweryfikuj wiarygodność podanych źródeł: czy rzeczywiście istnieją?
- Czy są one dostępne?
- Czy podają właściwych autorów?

Jeśli dostrzeżesz błąd sztucznej inteligencji, podziel się swoimi spostrzeżeniami z grupą oraz instruktorem.

Scenariusz: ChatGPT 4.0 vs NotebookLM

W analizie przypadku porównano dwa narzędzia przeznaczone do realizacji tego samego zadania:
„Generowanie definicji strategii innowacji z bibliografią”.

To samo polecenie wygenerowało dwa różne wyniki:

- ChatGPT 4.0: precyzyjna definicja, brak autentycznej bibliografii, zidentyfikowany jako 100% AI przez Originality.ai.
- NotebookLM (z autentycznymi plikami PDF): uzasadniona definicja opierająca się na rzeczywistych autorach (Porter, Janasz, Pomykalski, Freeman); 92% oceny oryginalnej.

Co to oznacza w praktyce?

- Bez solidnych podstaw, sztuczna inteligencja generuje płynne, lecz nieudokumentowane treści. Dysponując rzeczywistą bazą wiedzy, ten sam model wytwarza weryfikowalny tekst związany z konkretnymi publikacjami.

ChatGPT 4.0 a NotebookLM:

Źródła nie mają wyłącznie charakteru naukowego – to one kształtują odpowiedzialność za wyniki sztucznej inteligencji.

ChatGPT 4.0 (bez osadzenia):

- Płynny tekst, zwięzłe definicje, brak bibliografii
- Prawdopodobne przykłady przedsiębiorstw (Google, Tesla, Apple) nieposiadające żadnych „badań” w swoim dorobku.
- 100% wykryto jako wygenerowane przez sztuczną inteligencję przez Originality.ai

NotebookLM (z załadowanymi dokumentami PDF):

- Definicje oparte na teorii, odwołujące się do Portera, Janasza, Pomykalskiego i Freemana
- Prawdziwe tradycje badawcze: typologie innowacji, gospodarka sieciowa, kontekst małych i średnich przedsiębiorstw.
- Konkretna bibliografia w formacie APA, przygotowana do weryfikacji.
- Styl oceniony na 92% jako autentyczny przez Originality.ai

Źródła i prawa:

„Czy posiadam jakiekolwiek rzeczywiste źródło dla tego twierdzenia?”

Niezawodne zastosowanie sztucznej inteligencji to nie tylko fakty. To także kwestia autorstwa: kto jest twórcą oryginalnego pomysłu, czyje dzieło przytaczasz i w jaki sposób je przytaczasz.

Zazwyczaj można polegać na sztucznej inteligencji, jednak wyniki AI nie powinny być traktowane jako wiarygodne w następujących kwestiach:

- Dokładne statystyki, procenty, rozmiary próby
- Konkretne cytaty przypisywane wskazanym autorom
- Lata publikacji, numery stron, tytuły periodyków
- Definicje niszowych lub innowacyjnych koncepcji, które nie występują w literaturze głównego nurtu.
- Ogólne wyjaśnienia powszechnie znanych pojęć
- Edycja strukturalna (gramatyka, płynność, akapity)
- Podsumowanie tekstu, który sam dostarczyłeś sztucznej inteligencji.

Jak definiuje się „prawdziwe odniesienie” (i czym nie jest)?

Niezawodne zastosowanie sztucznej inteligencji to nie tylko fakty. To także kwestia autorstwa: kto jest twórcą oryginalnego pomysłu, czyje dzieło przytaczasz i w jaki sposób je przytaczasz.

Prawdziwe źródło to nie tylko nazwisko i rok. W przypadku cytatów generowanych przez sztuczną inteligencję źródło musi cechować:

- **Rzeczywistość**: publikacja rzeczywiście istnieje i jest dostępna w bazie danych.
- **Konkretność**: dosłowne dopasowanie żądania, a nie jedynie tematu
- **Dostępność**: możesz ją otworzyć i zapoznać się z odpowiednią sekcją.
- **Autorytatywność**: opublikowana przez rzetelne czasopismo, wydawcę lub instytucję
- **Ważne**: zgoda jest uzależniona od odbiorców oraz narzędzia.

Ktoś może zgodzić się przekazać ci pewne informacje, ale nie wyrazić zgody na ich udostępnienie narzędziu opartemu na sztucznej inteligencji lub na ich przechowywanie w systemie.

05

Halucynacje sztucznej inteligencji

(10 minut)

Czym są halucynacje sztucznej inteligencji?

Halucynacje sztucznej inteligencji to odpowiedzi generowane przez model, które brzmią wiarygodnie, lecz zawierają fałszywe, zmyślone lub wprowadzające w błąd informacje. Model „nie wie”, co jest prawdą – zazwyczaj przewiduje najbardziej prawdopodobny tekst, zamiast weryfikować rzeczywistość.

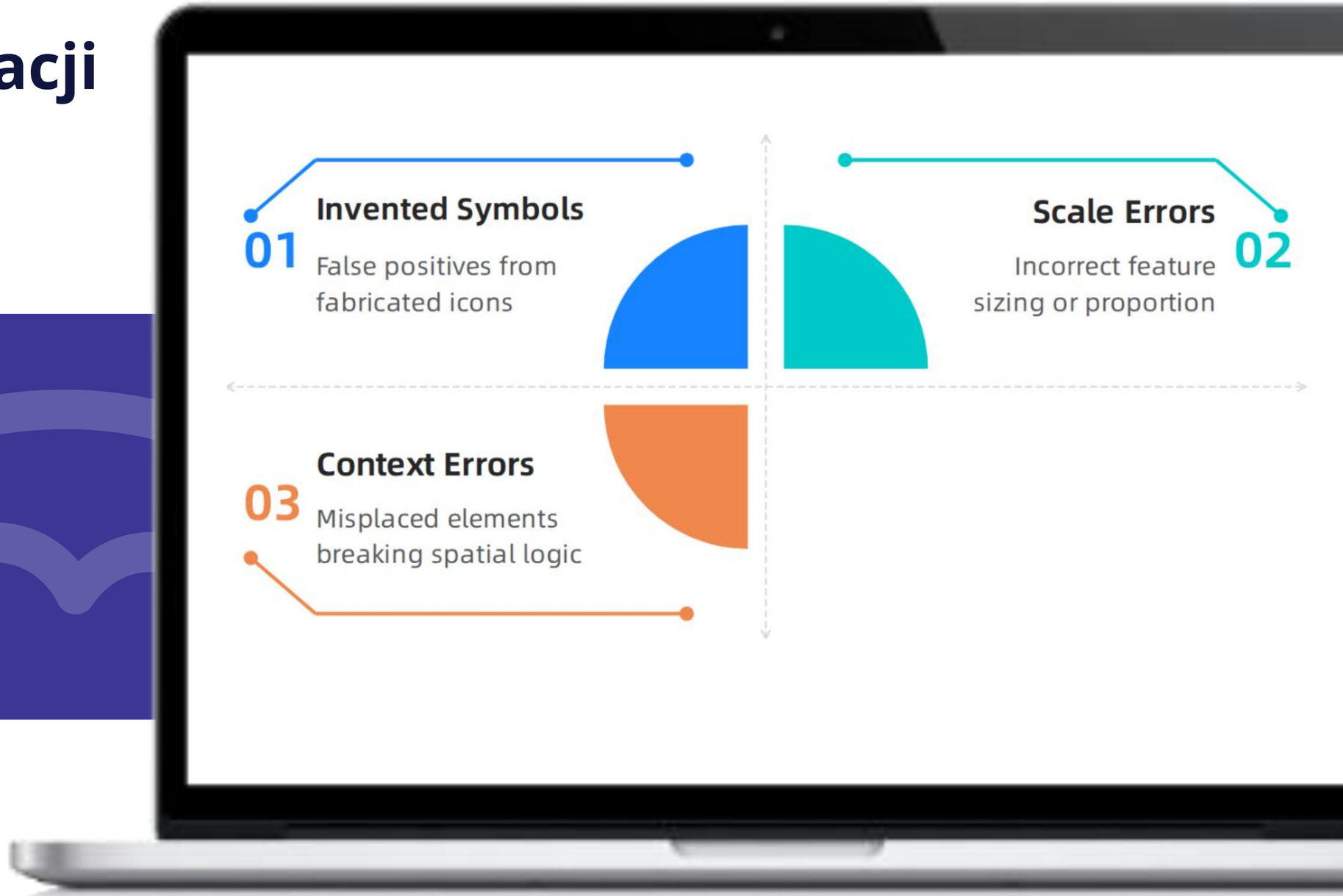
Halucynacje AI mają miejsce, gdy system tworzy treść, która wydaje się poprawna, lecz jest nieprawdziwa pod względem faktograficznym, na przykład zmyślony cytat, nieistniejące źródło, błędna data lub mylące wyjaśnienie. Nie są to halucynacje w sensie medycznym – to błędy w generowaniu informacji, a nie doświadczenia percepcyjne.

Mapa halucynacji sztucznej inteligencji

Jak identyfikować i naprawiać halucynacje sztucznej inteligencji w mapach symboli stopniowanych.

Aihallucinationreport.com.

https://aihallucinationreport.com/how-to-detect-and-fix-ai-hallucinations-in-graduated-symbol-maps/



Halucynacje sztucznej inteligencji

Statystyki dla 284 jednostek analizy. Różne kolory wskazują różne typy błędów.

Sun, Yujie, Sheng, Dongfang, Zhou, Zihan, Wu, Yifei. (2024). Halucynacja AI: w kierunku kompleksowej klasyfikacji zniekształconych informacji w treściach generowanych przez sztuczną inteligencję. Komunikacja w naukach humanistycznych i społecznych, 11. 10.1057/s41599-024-03811-x.

Rodzaje halucynacji sztucznej inteligencji

01

Halucynacje faktyczne – model przedstawia nieprawdziwe fakty, zdarzenia, osoby lub definicje.

02

Halucynacje źródłowe – sztuczna inteligencja tworzy fikcyjne źródła, publikacje, autorów, tytuły lub przypisuje sobie prawa do nieistniejących tekstów.

03

Halucynacje cytatowe – model przedstawia cytat, który w rzeczywistości nigdy nie został wypowiedziany lub błędnie przypisuje go danej osobie.

04

Halucynacje liczbowe – sztuczna inteligencja niepoprawnie interpretuje dane, myli procenty, daty, statystyki lub wartości tabelaryczne.

Rodzaje halucynacji sztucznej inteligencji

05

Halucynacje kontekstowe – model błędnie interpretuje pytanie lub wprowadza szczegóły, których użytkownik nie zamierzał uwzględnić.

06

Mieszane halucynacje – odpowiedź zawiera zarówno rzeczywiste fakty, jak i fikcyjne elementy, co utrudnia identyfikację błędu.

Przyczyny halucynacji sztucznej inteligencji

01

Model przewiduje wzorce, a nie prawdę – generuje najbardziej prawdopodobne słowa, zamiast weryfikować fakty.

02

Dane szkoleniowe mogą być niepełne lub błędne – jeśli zawierają luki, błędy lub sprzeczności, model jest w stanie je zrekonstruować.

03

Brak bieżących informacji – model może nie być świadomy najnowszych wydarzeń lub zmian.

04

Niejasny kontekst wskazówek – w przypadku niejednoznacznych lub skrótowych wskazówek model może błędnie interpretować intencje użytkownika.

Przyczyny halucynacji sztucznej inteligencji

05

Systemy nagradzają płynne odpowiedzi – modele są często trenowane w taki sposób, aby udzielały pewnych i płynnych odpowiedzi, nawet w sytuacjach, gdy występuje niepewność.

06

Brak wbudowanej funkcji weryfikacji faktów – bez dostępu do zewnętrznych źródeł model nie jest w stanie potwierdzić, czy jego odpowiedź jest prawdziwa.

Halucynacje pośrednie:
kiedy imiona i daty wydają się poprawne, ale w rzeczywistości takie nie są

Typowe halucynacje pośrednie:

- Prawdziwy autor + prawdziwe czasopismo + fikcyjny tytuł artykułu
- Prawdziwy dokument + nieprawidłowy rok, strona lub numer tomu
- Prawdziwa koncepcja + niewłaściwy autor („Porter stwierdził”, gdy był to Schumpeter)
- Rzeczywista statystyka + fikcyjny kontekst (prawidłowa liczba, błędne badanie)

Bezpieczniejsze opcje:

- **Wyszukaj**: otwórz Google Scholar i zweryfikuj, czy dany dokument istnieje.
- **Zastąp**: przytocz artykuł, który rzeczywiście przeczytałeś na ten sam temat
- **Zmiękcz**: zastosuj ogólne sformułowania, jeśli nie jesteś w stanie wskazać konkretnego źródła.
- **Usuń**: jeśli nie możesz zlokalizować źródła, usuń całe oświadczenie.

Sztuczna inteligencja może doświadczać halucynacji, nawet gdy imiona, daty i tytuły wydają się być poprawne. Schemat jest prawdopodobny, jednak związek z oryginalnym dziełem jest naruszony.

06

Weryfikacja treści

(10 minut)

Treści generowane przez sztuczną inteligencję mogą być użyteczne, jednak nie powinny być publikowane ani wykorzystywane bez uprzedniej weryfikacji.

Sprawdź twierdzenia, liczby, cytaty oraz źródła przed ich użyciem.

Potraktuj odpowiedź sztucznej inteligencji jako wstępny zarys, a nie jako ostateczną prawdę.

Udokumentuj, co zostało potwierdzone oraz na jakiej podstawie.

Krok 1:

Zidentyfikuj ryzyko oraz kluczowe roszczenia

Najpierw zidentyfikuj elementy odpowiedzi AI, które wymagają obowiązkowego sprawdzenia: fakty, daty, nazwy, liczby, cytaty, definicje oraz cytowane źródła. Im większy wpływ na zdrowie, prawo, edukację, finanse lub reputację, tym bardziej rygorystyczna powinna być weryfikacja.

Które elementy mogą spowodować szkody, jeśli są nieprawidłowe?

Czy sztuczna inteligencja dostarczyła autentyczne źródła, czy jedynie brzmi przekonująco?

Co w tej odpowiedzi stanowi fakt, a co jest interpretacją?

Krok 2:

Weryfikacja faktów i źródeł

Każde istotne stwierdzenie powinno być weryfikowane na podstawie wiarygodnych źródeł zewnętrznych, najlepiej pierwotnych: artykułów naukowych, raportów, stron internetowych instytucji publicznych oraz renomowanych publikacji branżowych. Nigdy nie zakładaj, że źródło cytowane przez sztuczną inteligencję rzeczywiście istnieje — zweryfikuj autora, tytuł, rok, dostępność oraz to, czy źródło rzeczywiście zawiera podane informacje.

Potwierdź informację w co najmniej dwóch niezależnych źródłach.

Otwórz oryginalne źródło, a nie jedynie streszczenie lub powtórne opublikowanie.

Sprawdź datę publikacji oraz aktualność informacji.

Krok 3:

Oceń jakość, język oraz kontekst

Po weryfikacji faktów oceń, czy tekst jest logiczny, klarowny, zrównoważony i adekwatny do zamierzonego celu. Sztuczna inteligencja potrafi generować treści, które są formalnie poprawne, lecz słabo osadzone w kontekście, nadmiernie pewne siebie, stronnicze lub stylistycznie nieodpowiednie.

W tekście fakty są pomieszane z opiniami.

Zawiera stronniczość, nadmierne uproszczenia oraz zbyt daleko idące uogólnienia.

Ton i styl są dostosowane do odbiorców oraz celu publikacji.

Edytuj, podejmuj decyzje i dokumentuj

Na koniec popraw błędy, usuń niejasne fragmenty i zdecyduj, czy treść jest odpowiednia do użytku. Dobrą praktyką jest odnotowanie, co zostało zmienione, co odrzucone oraz które źródła potwierdziły ostateczną wersję tekstu. Decyzja końcowa:

Publikuj

Jeśli treść została zweryfikowana i skorygowana.

Wstrzymaj się

Jeśli niektóre informacje pozostają niejasne.

Odrzuć

Jeśli treść zawiera zbyt wiele błędów lub brakuje weryfikacji źródła.

07

Analiza danych z zastosowaniem sztucznej inteligencji

(10 minut)

Gdy analizujesz dane przy użyciu sztucznej inteligencji, walidacja staje się kluczowym zadaniem.

Gdy zwrócisz się do sztucznej inteligencji o analizę danych, system stworzy spójną narrację. W zależności od polecenia i zastosowanego narzędzia, rezultat może również:

- Pominąć metodologię (brak odniesienia do wielkości próby, metody lub założeń)
- Uśrednić wartości
- Uogólnić na podstawie ograniczonej próby (np. n=10 respondentów)
- Wyciągać konkretne wnioski (gdy dane potwierdzają jedynie wnioski wstępne)
- Błąd potwierdzenia: sztuczna inteligencja może powtarzać treść komunikatu.

Gdy analizujesz dane przy użyciu sztucznej inteligencji, walidacja staje się kluczowym zadaniem.

Dlaczego walidacja analizy jest istotna:

- Wykrywasz błędy zanim trafią do raportu lub publikacji.
- Potrafisz rozróżnić solidne ustalenia od przekazu informacyjnego, który wydaje się wiarygodny.
- Unikasz podejmowania decyzji na podstawie próbek, które nie są reprezentatywne.
- Zachowujesz autorstwo analizy, a nie modelu.

Drabina analizy SAFE: wybierz najbardziej weryfikowalny poziom

Analizując dane przy użyciu sztucznej inteligencji, nie rozpoczynaj od najwyższego poziomu. Wybierz najniższy poziom, który nadal odpowiada na Twoje pytanie.

Zasada zatrzymania: Jeśli nie jesteś w stanie usunąć wrażliwych lub poufnych elementów → nie przesyłaj.

POZIOM 1

(Najbardziej wiarygodny): Zadaj pytanie ogólne (bez danych osobowych).

„Wyjaśnij, w jakich okolicznościach test t jest stosowny, a kiedy należy zastosować test U Manna-Whitneya."

POZIOM 2

Użyj anonimowego podsumowania (role, bez identyfikatorów).

„Średnia = 4,2, SD = 1,1, n = 120 – co to implikuje w kontekście trendu?"

POZIOM 3

Używaj ograniczonych, zanonimizowanych danych wyłącznie w przypadku konieczności.

Przykład: „Oto zestaw danych. Oblicz X, a następnie ja porównam to z moimi obliczeniami."

Wnioski z ograniczonych danych

Niepewny rezultat sztucznej inteligencji (realistyczny):

„Sztuczna inteligencja przeprowadziła analizę sentymentu na podstawie 200 tweetów dotyczących naszej marki. 67% z nich miało charakter pozytywny. W związku z tym sentyment wobec marki jest zdecydowanie pozytywny.”

Co jest niepewne?

- „67%” bez jakiejkolwiek metodologii, modelu i walidacji
- 200 tweetów to niewielki, prawdopodobnie subiektywny fragment.
- „zdecydowanie pozytywny” to przesada w odniesieniu do tego, co wskazuje liczba.

AI Skills, Powered by Values

values

Śledź naszą wędrówkę

www.valuesai.eu

Co-funded by the European Union

Współfinansowane przez Unię Europejską. Wyrażone poglądy i opinie są wyłącznie poglądami i opiniami autora lub autorów i niekoniecznie odzwierciedlają stanowisko Unii Europejskiej ani Fundacji Rozwoju Systemu Edukacji. Ani Unia Europejska, ani instytucja przyznająca grant nie ponoszą za nie odpowiedzialności.