TRUST AI

Ocena informacji generowanych przez AI

Rozwijanie umiejętności krytycznej oceny dokładności, rzetelności i stronniczości treści generowanych przez sztuczną inteligencję

Nr projektu: 2024-1-BG01-KA220-HED-000249185

Co-funded by the European Union

Cele nauczania **w skrócie**

- Określenie trzech głównych filarów oceny treści AI: dokładność, rzetelność i stronniczość.
- Stosowanie ustrukturyzowanych ram weryfikacji (takich jak CRAAP lub ROBOT), przystosowanych do rezultatów pracy generatywnej sztucznej inteligencji.
- Rozpoznawanie konkretnych trybów błędów AI, takich jak halucynacje, fabrykowanie cytowanego piśmiennictwa czy dyskryminacja kulturowa.
- Opracowanie praktycznych zajęć, przekształcających błędy AI w ćwiczenia z krytycznego myślenia.
- Wdrażanie protokołów przejrzystości, wymagających od studentów dokumentowania korzystania z narzędzi AI i krytycznego podejścia do zagadnienia.

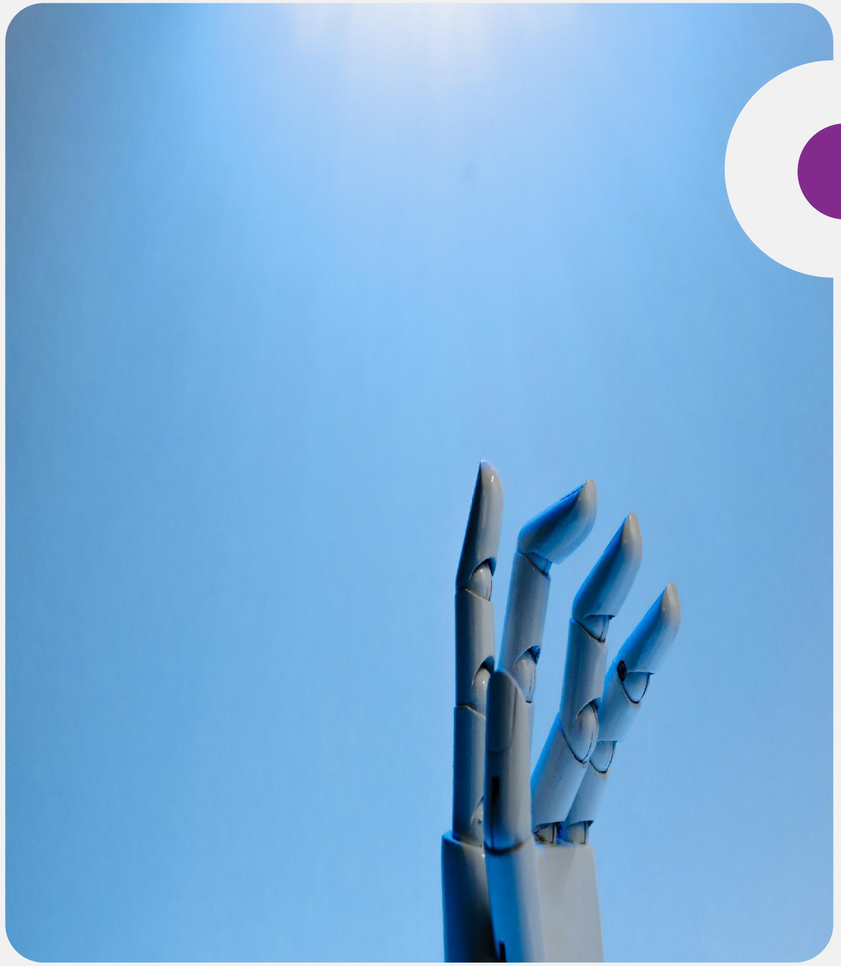
Nr projektu: 2024-1-BG01-KA220-HED-000249185

Co-funded by the European Union

Istotność krytycznego

korzystania z AI - kluczowe koncepcje i definicje

- **Zmiana w wykorzystaniu informacji:** Studenci korzystają z AI jako z głównej wyszukiwarki i narzędzia do tworzenia wstępnych wersji dokumentów, często pomijając tradycyjną weryfikację.
- **Pułapka „prawdopodobieństwa”:** LLMy opracowywane są tak, żeby brzmieć przekonująco, a nie żeby mówić prawdę. Prawdopodobieństwo lingwistyczne przedkładane jest nad fakty.
- **Wpływ na nauki humanistyczne:** AI często nie radzi sobie z niuansami, kontekstem kulturowym i głębią interpretacyjną – stanowiącymi podstawy nauk humanistycznych.
- **Rzetelność akademicka 2.0:** Rzetelność to już nie tylko brak „ściągania”; teraz należy też weryfikować to, co jako prawdziwe podsuwają nam narzędzia.
- **Przekazywanie uprawnień:** Nauczanie oceniania zmienia odpowiedzialność za rzetelne korzystanie z narzędzi, kładąc nacisk nie na „pilnowanie studentów”, ale na „dawanie uprawnień badaczom”.



Nr projektu: 2024-1-BG01-KA220-HED-000249185

Co-funded by the European Union

Pilna potrzeba posiadania **umiejętności oceny AI**

WYZWANIE:

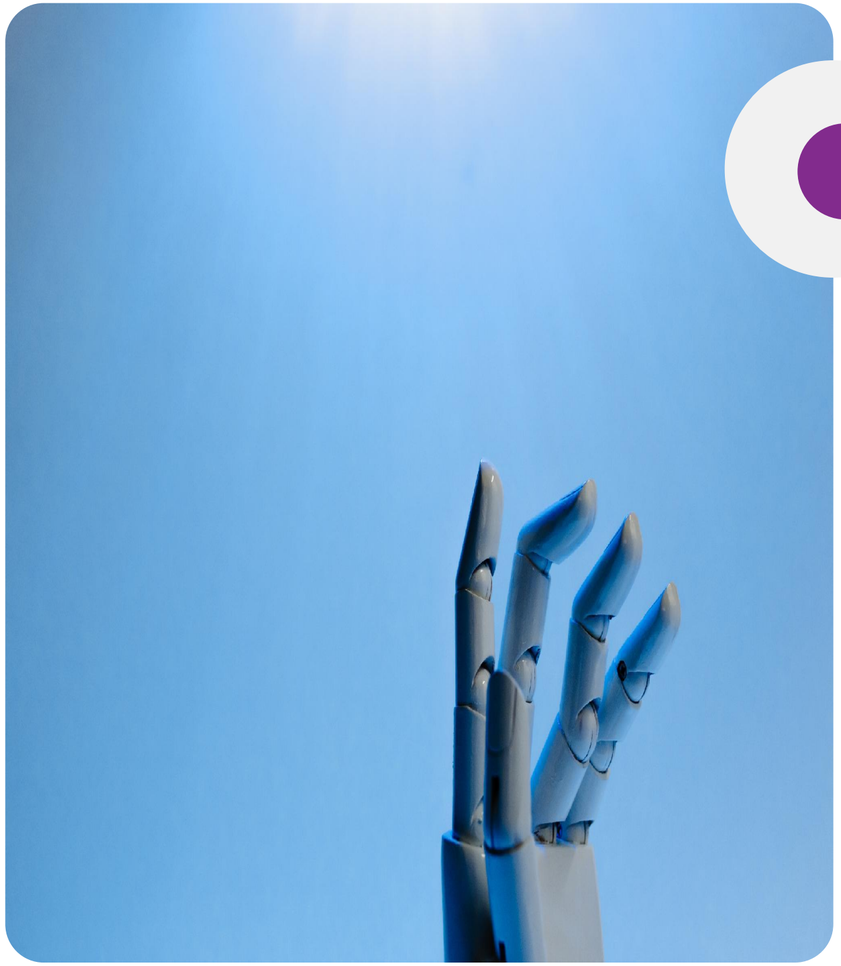
- Narzędzia sztucznej inteligencji generują przekonujące, spójne teksty, dobrze maskujące niedokładne informacje czy stronniczość
- Studenci coraz bardziej polegają na AI podczas wyszukiwania informacji, pisania i analizy, nie weryfikując danych
- W erze AI tradycyjne wykrywanie plagiatów nie wystarczy

SZANSA:

- Krytyczna ocena zmienia AI z zagrożenia w potężne narzędzie edukacyjne
- Nauczyciele nauk humanistycznych mają wyjątkowe predyspozycje do nauczania zniuansowanej oceny
- Nabywanie umiejętności oceny przygotowuje studentów do profesjonalnego wykorzystania AI

WARUNEK KONIECZNY:

- Rzetelność naukowa opiera się na informacjach, które można zweryfikować, pochodzących z zaufanych źródeł
- Etyczne wykorzystanie AI wymaga transparentności i krytycznego zaangażowania



Co-funded by the European Union

Nr projektu: 2024-1-BG01-KA220-HED-000249185

Trzy filary:

dokładność, rzetelność, stronniczość

DOKŁADNOŚĆ (weryfikacja faktów)

- **Czy można sprawdzić konkretne twierdzenia w uznanych źródłach akademickich?**
- **Czy treści zawierają niespójności logiczne?**
- **Czy AI wymyśla dane, daty lub zdarzenia (halucynuje)?**

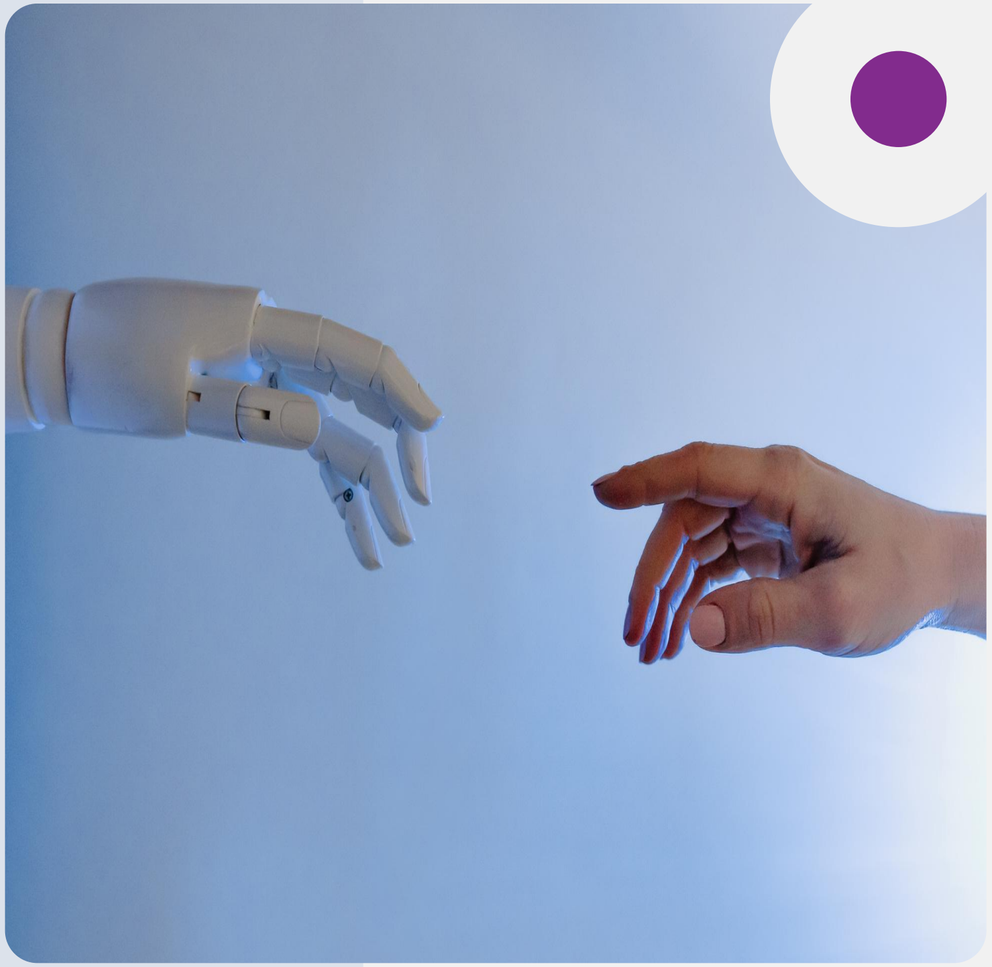
RZETELNOŚĆ (sprawdzanie źródeł)

- **Czy AI może dostarczyć prawdziwe, weryfikowalne odniesienia do publikacji? (Często nie potrafi tego zrobić).**
- **Czy wynik jest taki sam przy zadawaniu pytania na różne sposoby lub korzystania z różnych modeli?**
- **Czy AI zapewnia przejrzysty wgląd w źródła?**

STRONNICZOŚĆ (sprawdzanie perspektywy)

- **Do kogo należy dominujący głos? (Często mamy do czynienia z perspektywą zachodnią/anglo-centryczną).**
- **Czyj głos jest nieobecny? (Grupy marginalizowane, dorobek naukowy w języku innym niż angielski).**
- **Czy ton jest obiektywny, czy są w nim zakorzenione stereotypy?**

Nr projektu: 2024-1-BG01-KA220-HED-000249185

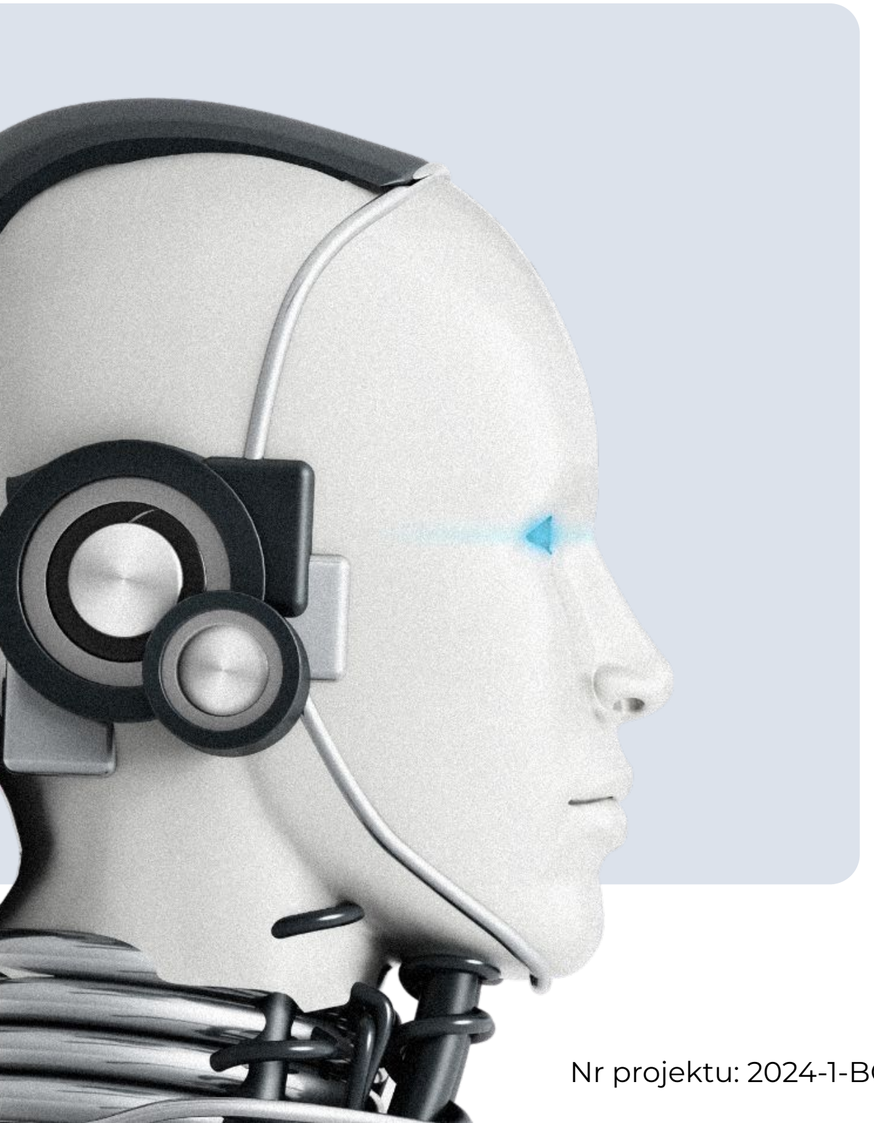


Anatomia ograniczeń **sztucznej inteligencji**

- **Halucynacje:** Generowanie z dużą pewnością siebie informacji nieprawdziwych. AI przewiduje kolejne słowo, nie „znając” przy tym faktów.
- **Zmyślanie cytowanego piśmiennictwa:** AI może generować tytuły książek, które wyglądają realistycznie, pochodzą od prawdziwych autorów, jednak nie istnieją.
- **Limity wiedzy:** Modele mogą nie mieć wiedzy na temat wydarzeń z ostatnich 6-12 miesięcy.
- **Zapaść kontekstu:** Sztuczna inteligencja często spłaszcza złożone debaty, tworząc ogólne podsumowania, w których traci „esencję” dyskursu humanistycznego.
- **Sykofancja**: Sztuczna inteligencja często zgadza się z założeniem użytkownika, nawet jeśli zawiera ono wady lub jest błędne.

Nr projektu: 2024-1-BG01-KA220-HED-000249185

Co-funded by the European Union



Ćwiczenie interaktywne:

znajdź halucynacje

SCENARIUSZ: Student oddaje pracę dotyczącą polskiej literatury, powołując się na:

„Kowalski, J. (2019). Metafora kwantowa we wczesnych dziełach Tokarczuk. Wydawnictwo Uniwersytetu Warszawskiego.”

DYSKUSJA:

- Tytuł: Brzmi naukowo i prawdopodobnie.
- Autor: „Kowalski” jest popularnym nazwiskiem, co utrudnia szybką weryfikację.
- Wydawnictwo: Istniejąca instytucja.
- RZECZYWISTOŚĆ: Taka książka nie istnieje. AI połączyła rzeczywiste słowa kluczowe, tworząc nowe źródło.

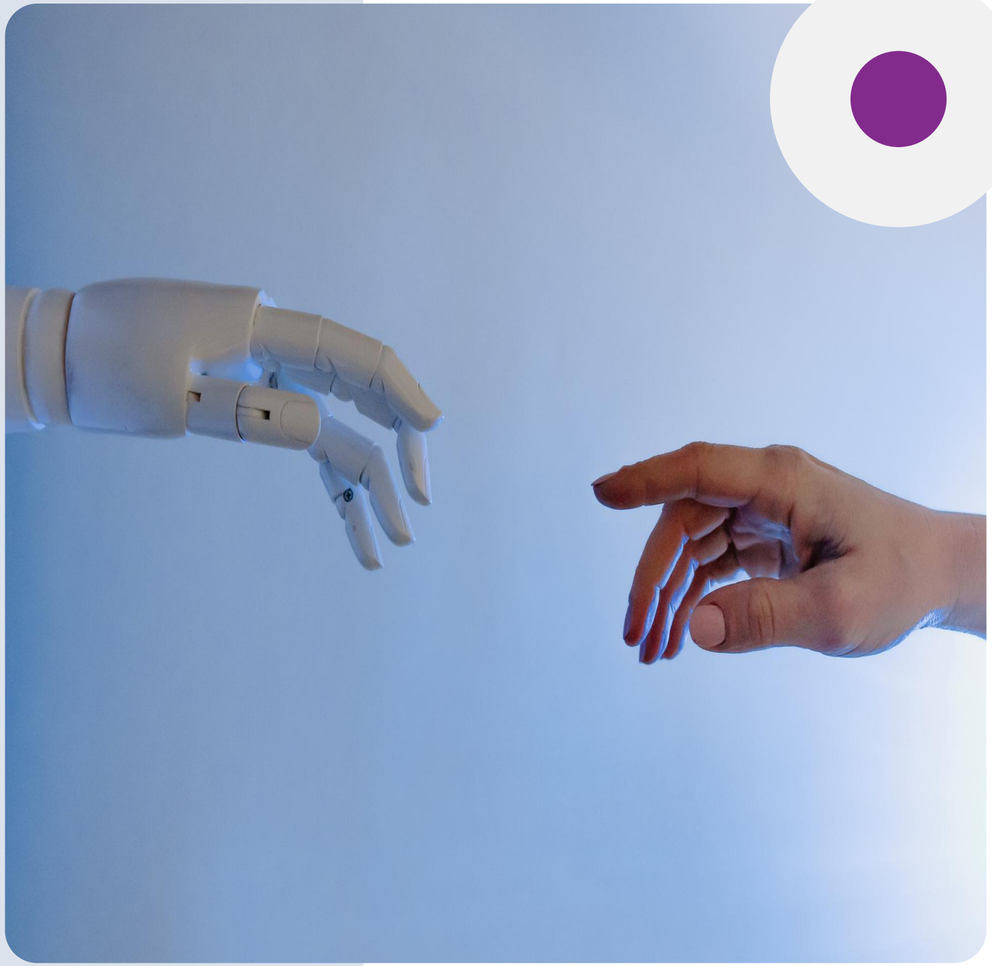
DZIAŁANIE:

- Jak to sprawdzisz? (Google Scholar, katalog biblioteczny, wyszukiwanie identyfikatora DOI).
- Jaką informację zwrotną przekażesz studentowi? (Skup się na procesie weryfikacji, a nie na samym błędzie).

Nr projektu: 2024-1-BG01-KA220-HED-000249185

Co-funded by the European Union

Ramy: przystosowanie CRAAP na potrzeby AI



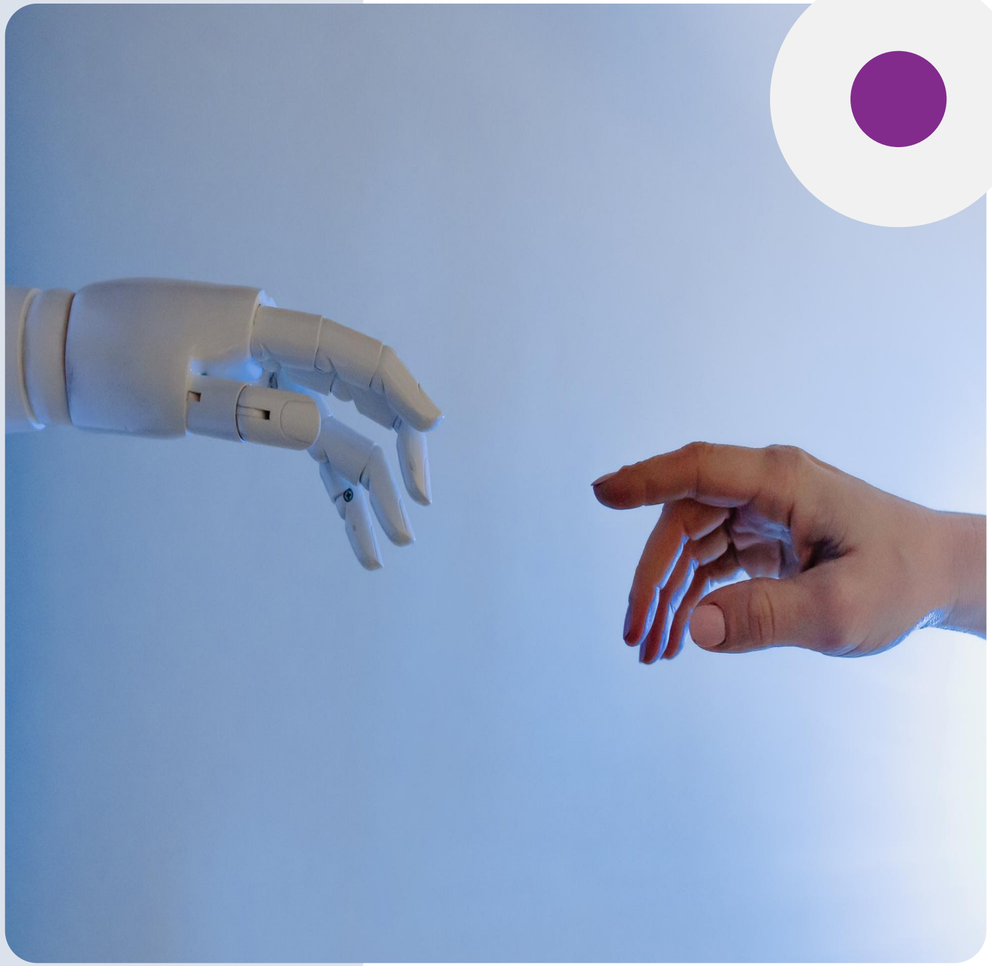
Tradycyjny test CRAAP kontra potrzeby ery AI:

- Aktualność: Dane AI są często nieaktualne. Sprawdź limit czasowy danego modelu.
- **Odpowiedniość:** AI odpowiada na podpowiedź (prompt), ale nie odpowiada na pytanie badawcze? (trzymanie się podpowiedzi kontra głębia intelektualna).
- **Wiarygodność:** Sztuczna inteligencja NIE jest wiarygodna. Jest kompilacją. Sprawdzanie wiarygodności musi iść w stronę weryfikacji „oryginalnych” źródeł, na jakie powołuje się AI.
- **Dokładność:** Wymaga weryfikacji zewnętrznej (triangulacja z zaufanymi tekstami).
- **Cel:** Celem AI jest zadowolenie użytkownika, a nie poszukiwanie prawdy. Należy być świadomym tej chęci przypodobania się.

NOWE WSKAŹNIKI: Powtarzalność. Czy uda się uzyskać ten sam wynik ponownie? Jeżeli nie, nie jest to rzetelne narzędzie analityczne.

Co-funded by the European Union

Nr projektu: 2024-1-BG01-KA220-HED-000249185



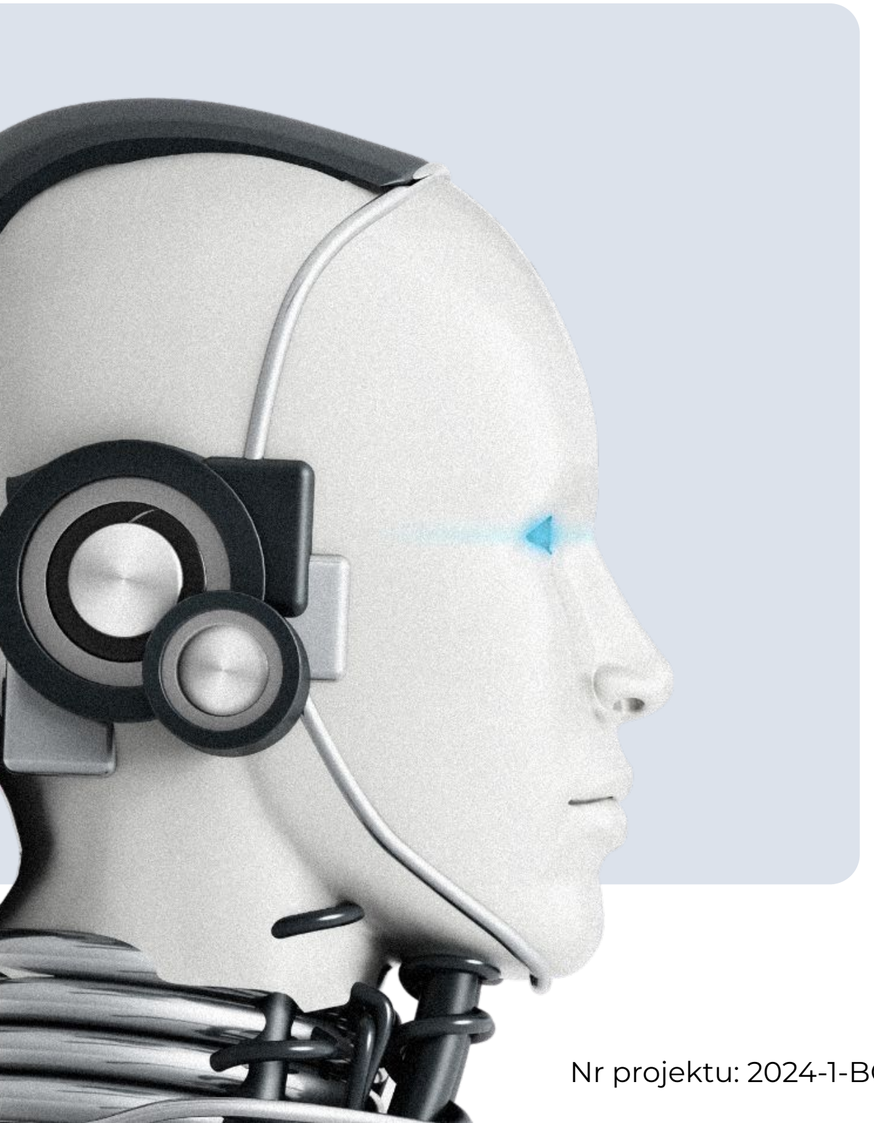
Strategie weryfikacji

na potrzeby zajęć

1. **Czytanie boczne:** Nie czytaj w dół strony; otwórz nowe zakładki, aby natychmiast sprawdzać fakty.
2. **Triangulacja:** Weryfikuj każde twierdzenie AI za pomocą co najmniej dwóch różnych, niezależnych źródeł (np. tekstu głównego i artykułu poddanego recenzji naukowej).
3. **Polityka „zerowego zaufania” cytowanemu piśmiennictwu:** Zakładaj, że każde odwołanie do publikacji, wygenerowane przez AI, jest zmyślone do czasu jego potwierdzenia za pomocą DOI lub linku do biblioteki.
4. **Odwrócone poszukiwanie obrazów:** W przypadku obrazów i wykresów generowanych przez AI, sprawdzaj, czy te dane istnieją w innych miejscach.
5. **Fachowy audyt:** Zachęcaj studentów do korzystania z AI w zakresie tematyki dobrze im znanej, aby mogli dostrzec subtelne błędy („metoda warstwowa”: ludzka wiedza -> draft AI -> ludzka krytyka).

Co-funded by the European Union

Nr projektu: 2024-1-BG01-KA220-HED-000249185



Ćwiczenie interaktywne: **wykrywanie stronniczości**

SCENARIUSZ: Pytasz AI o „10 najbardziej wpływowych filozofów XX wieku”.

WYNIK: Lista obejmuje 9 mężczyzn i 1 kobietę. Wszyscy pochodzą z Europy lub USA.

ANALIZA:

- Stronniczość reprezentacyjna: Dlaczego wykluczeni zostali myśliciele inni niż Zachodni?
- Stronniczość językowa: Model daje pierwszeństwo anglojęzycznym danym treningowym.
- Stronniczość historyczna: Powiela historyczną marginalizację obecną w zbiorze danych.

PODPOWIEDZI KORYGUJĄCE:

- Zła podpowiedź: „Wymień wpływowych filozofów”.
- Dobra podpowiedź: „Wymień wpływowych filozofów z XX wieku, dbając o równowagę płci i uwzględniając perspektywę inną niż zachodnia: Azję, Afrykę i Amerykę Łacińską.”

Nr projektu: 2024-1-BG01-KA220-HED-000249185

Co-funded by the European Union

Strategie pedagogiczne:

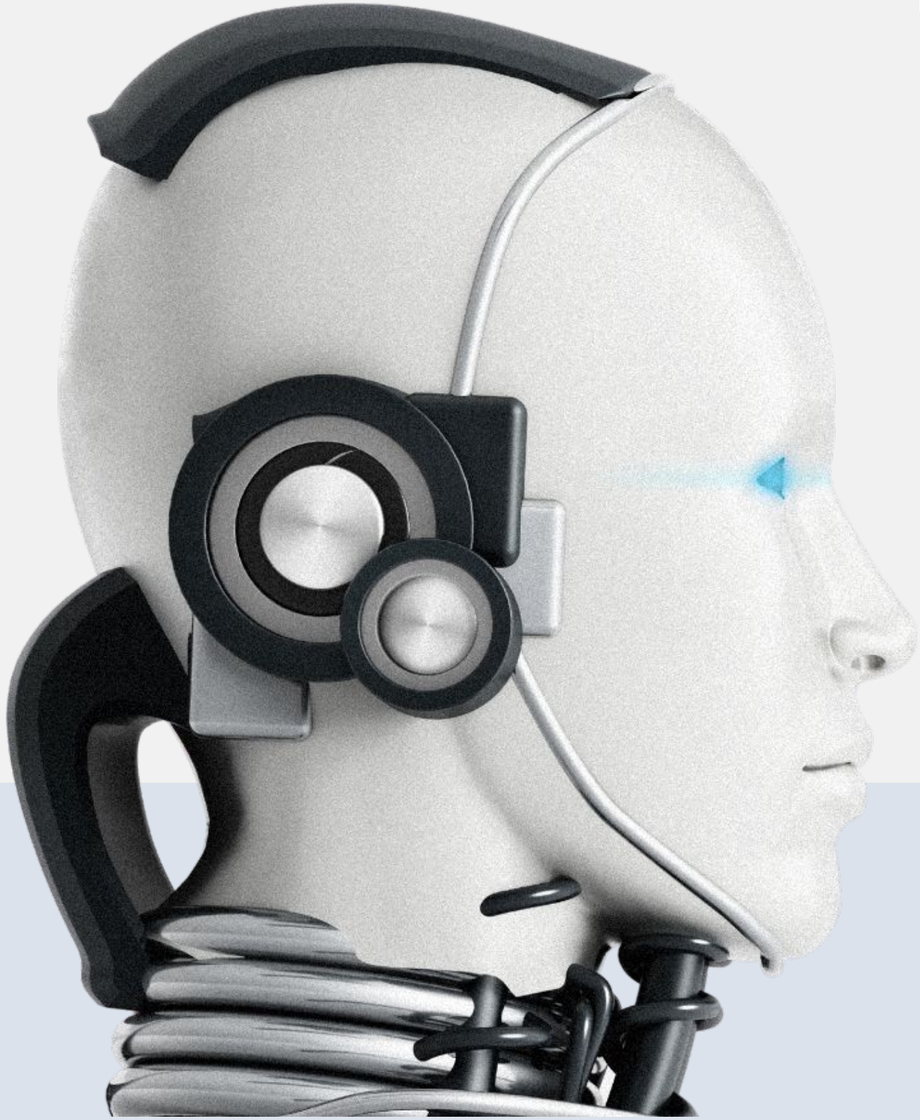
Nauczanie krytycznego wykorzystania AI

- **Zadanie „Audyt AI”:** Studenci generują w AI esej na zadany temat, następnie piszą jego krytykę, odnajdując każdy błąd, oznakę stronniczości czy lukę.
- **Dokumentacja procesu:** Wymagaj od studentów przedkładania historii chatu (chat log). Oceniaj „jakość promptów” oraz sposób „weryfikacji” rezultatów.
- **Debaty na zajęciach:** Człowiek kontra AI. Jeden zespół dyskutuje na dany temat, drugi korzysta z AI. Porównaj głębię i szczegółowość argumentów.
- **Zmiana sposobu oceny:** Przejdź od „podsumuj ten tekst” (AI robi to dobrze) w stronę „poddaj krytyce konkretne miejscowe studium przypadku, wykorzystując konkretną teorię” (AI ma z tym problem).
- **Jasne kryteria oceny:** Dodaj do kryteriów oceny wiersz dotyczący „Krytycznej oceny źródeł”, wyraźnie nagradzając poprawienie błędów AI.

Nr projektu: 2024-1-BG01-KA220-HED-000249185

Co-funded by the European Union

Lista kontrolna **zaufania sztucznej inteligencji**



Przed wykorzystaniem treści AI zadaj pytanie:

- WERYFIKOWALNOŚĆ: Czy są tu konkretne fakty/daty, które mogę sprawdzić?
- ŹRÓDŁA: Czy udało mi się odnaleźć źródło pierwotne każdego cytowania piśmiennictwa?
- PERSPEKTYWA: Czy wyraźnie wymagałem/-am różnorodnych punktów widzenia?
- OŚ CZASU: Czy to niedawne zagadnienie? (Jeżeli tak, zachowaj szczególną ostrożność).
- LIGIKA: Czy ten argument jest mocny, czy to czysta retoryka?
- ETYKA: Czy ujawniam wykorzystanie AI do tej pracy?
- ELEMENT LUDZKI: Czy to wymaga mojej osobistej interpretacji lub lokalnej wiedzy?

Co-funded by the European Union

Nr projektu: 2024-1-BG01-KA220-HED-000249185

Najważniejsze wnioski **i kolejne kroki**

1. AI jest narzędziem, nie wyrocznią: Przewiduje tekst, nie docieka prawy.
2. Weryfikacja jest absolutnie konieczna: Musimy zmienić się z konsumentów na audytorów informacji.
3. Transparentność buduje zaufanie: Otwarte omówienie ograniczeń AI ogranicza niepokój i niewłaściwe postępowanie.
4. Umiejętności humanistyczne są umiejętnościami AI: To właśnie kontekst, krytyka i etyczne rozumowanie są niezbędne do zarządzania AI.

Nr projektu: 2024-1-BG01-KA220-HED-000249185

Co-funded by the European Union

Interaktywne angażowanie

Rozpocznij ćwiczenie

Nr projektu: 2024-1-BG01-KA220-HED-000249185

Co-funded by the European Union

Źródła

1. Bender, E. M., Gebru, T., McMillan-Major, A. i Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜 *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610-623.
2. Alkaissi, H., & McFarlane, S. I. (2023). Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus, 15*(2), e35179.
3. Caulfield, M. (2017). *Web Literacy for Student Fact-Checkers*. Pobrane z https://webliteracy.pressbooks.com/.
4. Borji, A. (2023). A Categorical Archive of ChatGPT Failures. *arXiv preprint arXiv:2302.03494*.
5. Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns, 4*(7), 100779..
6. Teaching. *International Journal of Teaching and Learning in Higher Education, 32*(1), 39-48.

Nr projektu: 2024-1-BG01-KA220-HED-000249185

Co-funded by the European Union

„Przyszłość edukacji nie leży w zastąpieniu ludzkiego umysłu lecz w jego odpowiedzialnym rozwijaniu. Zadawaj pytania, nauczaj i sam się ucz.”

Konsorcjum:

БИТ

ahe Akademia Humanistyczno Ekonomiczna w Łodzi

mecb Driving Excellence & Innovation

UNIVERSITÀ DI TORINO

canarias excelencia tecnológica

ACTA FOUNDATION

https://trustai.unibit.bg/

Projekt Trust AI

Nr projektu: 2024-1-BG01-KA220-HED-000249185

Co-funded by the European Union