

Architektura Krytycyzmu

Strategie krytycznego myślenia w interakcji z GenAI i LLM-ami

mgr Ewelina Gajewska, mgr Katarzyna Kwiatkowska



Spis Treści

1. Wprowadzenie: czym jest krytyczne myślenie?
2. Metoda sokratejska.
3. Analiza argumentów.
4. Warsztat z LLM-ami: analiza przypadków.

Wprowadzenie

Czym jest krytyczne myślenie i dlaczego jest potrzebne?

| Zmiana Paradygmatu

Potoczne Rozumienie

Krytyczne myślenie kojarzone jest z akademicką debatą, "niezgadzaniem się" dla zasady lub poszukiwaniem argumentów na potwierdzenie z góry założonej tezy (tzw. myślenie prawnicze).

Dzisiejsza Konieczność

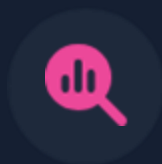
Dziś to Epistemiczna Czujność. To systematyczny proces weryfikacji i dekonstrukcji założeń. To umiejętność zawieszenia osądu do momentu znalezienia bezwzględnych dowodów.

Dlaczego LLM-y nie są narzędziami do krytycznego myślenia



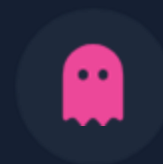
Potakiwanie

Model zgadza się z błędnymi założeniami użytkownika, by go zadowolić (efekt treningu RLHF).



Efekt potwierdzenia

AI chętniej generuje dane potwierdzające hipotezę z promptu, ignorując kontrargumenty.



Halucynacje

Generowanie wysoce prawdopodobnych statystycznie, ale merytorycznie fałszywych faktów i cytatów.



Złudzenie pewności

Język modelu brzmi zawsze autorytarnie, nawet gdy prawdopodobieństwo prawdy to 5%.

! Pułapka Potakiwacza

Modele takie jak ChatGPT są trenowane metodą **uczenia ze wzmocnieniem z informacją zwrotną**. Użytkownicy preferują odpowiedzi bardzo uprzejme i unikające konfliktów.

Wywołuje to zjawisko potakiwania, polegające na tym, że jeśli w prompcie zawrzesz własną opinię lub błąd, model z dużym prawdopodobieństwem zgodzi się z Tobą i dopisze do tego argumenty potwierdzające tę opinię, zamiast Cię poprawić.

⚠ Model staje się lustrem Twoich własnych uprzedzeń.



Jak traktujesz maszynę?

AI jako Wyrocznia

Błąd: LLM jako "lepszy Google". Prośenie o gotowe fakty, definicje i ostateczne wnioski.

Prowadzi to do ślepego zaufania. Jeśli model wygeneruje halucynację, użytkownik bezkrytycznie implementuje ją do swojej pracy.

AI jako Sparring Partner

Alternatywa: LLM jako asystent. Prośenie o stworzenie struktury, analizę argumentów, szukanie luk logicznych lub testowanie naszych hipotez.

Ostateczna decyzja i "myślenie" zawsze pozostaje po stronie ludzkiego umysłu.

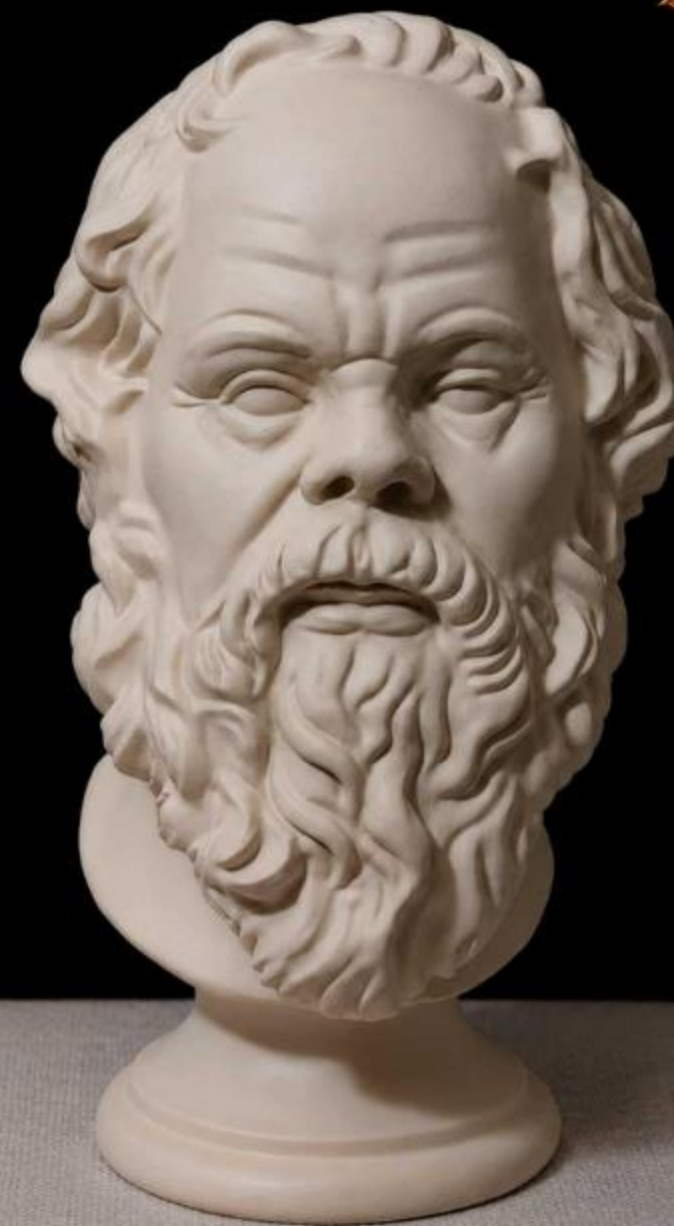
Metoda 1

Metoda Sokratejska: Geneza Metody

Starożytne Ateny, V wiek p.n.e.

Sokrates uważał, że uświadomienie sobie własnej niewiedzy to najwyższa forma mądrości. Jego metoda polega więc na krzyżowym ogniu pytań, mająca na celu obnażenie luk i sprzeczności logicznych.

Zamiast atakować tezę rozmówcy wprost, Sokrates zmuszał go do jej samodzielnego rozebrania za pomocą logicznych pytań.



| Typy Pytań Sokratejskich

-  **O precyzję** "Co dokładnie masz na myśli mówiąc [X]?"
-  **O założenia** "Zakładasz tutaj, że Y jest prawdą. Dlaczego?"
-  **O dowody** "Na jakich danych lub badaniach opierasz ten wniosek?"
-  **O perspektywę** "Jak wytłumaczyłby to ktoś o przeciwnych poglądach?"
-  **O konsekwencje** "Jeśli to prawda, to co to oznacza w dłuższej perspektywie?"
-  **Pytania o pytanie** "Dlaczego w ogóle uważasz to pytanie za istotne?"

Kontekst I: Zwykła debata

Atak Dogmatyczny

A: "Globalne ocieplenie to wymysł korporacji!"

B: "Jesteś ignorantem, nie znasz się na fizyce!
Wszyscy naukowcy mówią inaczej."

Efekt: Obronność, konflikt, brak refleksji.

Podejście Sokratejskie

A: "Globalne ocieplenie to wymysł korporacji!"

B: "To mocna teza. Jakie konkretnie anomalie w
pomiarach rdzeni lodowych skłaniają Cię do
odrzucenia oficjalnego konsensusu?"

*Efekt: Przeniesienie ciężaru dowodu, weryfikacja
wiedzy.*

| Kontekst II: Prompting LLM

Zjawisko Potakiwania

Prompt: "Wyjaśnij mi, dlaczego dieta ketogeniczna jest najlepsza dla zdrowia."

AI: Model potulnie wymyśli 10 powodów, pomijając ryzyka dla wątroby. AI dąży przede wszystkim do zadowolenia użytkownika.

Krytyczne Wymuszanie Balansu

Prompt: "Przedstaw biologiczne zalety diety keto. Następnie wciel się w surowego krytyka i znajdź 3 największe luki we własnym wywodzie."

AI: Model zostaje "zmuszony" do autorefleksji i podania holistycznej analizy.

Metoda 2

Anatomia Argumentu

Sztuka logicznego starcia i mapowania myśli

Elementy Składowe Argumentu



Główna Teza (Wniosek)

To, do czego próbujesz mnie przekonać. Główny punkt, wniosek końcowy. Np. *"Musimy wdrożyć AI w naszej firmie."*



Przesłanki (Dowody)

Fakty, dane lub obiektywne stwierdzenia, które mają wspierać konkluzję. Np. *"Konkurencja używa AI i rośnie o 20% szybciej."*



Wnioskowanie

Logiczny "klej". Relacja wiążąca przesłanki z konkluzją.

Teza vs. Argument

Sama Teza (Opinia)

"Sztuczna Inteligencja zniszczy świat i zabierze nam wszystkim pracę. Powinniśmy tego zakazać."

Diagnoza: Brak dowodów. Opiera się na emocjach i lęku. Nie można z tym dyskutować logicznie, bo nie ma struktury do obalenia.

Pełnoprawny Argument

"AI wpłynie na rynek pracy [TEZA], ponieważ potrafi zautomatyzować 40% powtarzalnych zadań biurowych [PRZESŁANKA 1], co jasno pokazują dane rynkowe z 2025 roku [PRZESŁANKA 2]."

Diagnoza: Mamy strukturę. Możemy teraz zweryfikować, czy dane z 2025 roku są prawdziwe (atak na Przesłankę 2).

Moduł 1



Rekonstrukcja i Mapowanie Argumentów

Jak rozłożyć tekst na części?

1. Znajdź Główną Tezę

Czego autor chce ode mnie? Szukaj słów: *zatem, więc, dlatego, musimy...*

2. Szukaj Przesłanek

Dlaczego mam w to uwierzyć? Szukaj słów: *ponieważ, bo, wynika to z, faktem jest...*

3. Usuń Szum

Wykreśl anegdoty, wyzwiska, emocjonalne przymiotniki. Zostaw czyste, nagie twierdzenia.

4. Narysuj Relację

Czy Przesłanka A bezpośrednio wspiera Konkluzję C? (Zrób strzałkę).

Ćwiczenie 1: Anatomia



TWÓJ RUCH

Przeczytaj poniższy tekst. Zidentyfikuj główną tezę oraz przesłankę, która ją wspiera:

"Wszyscy wiemy, że uczniowie są przemęczeni. Badania z zeszłego miesiąca pokazały, że 70% nastolatków śpi mniej niż 6 godzin dziennie z powodu zadań domowych. Dlatego też ministerstwo edukacji powinno całkowicie zakazać zadawania prac do domu na weekendy."

Moduł 2

Fallacie (Błędy Logiczne)

Kiedy Twój mózg idzie na skróty

Najczęściej Spotykane Błędy Logiczne



Ad Hominem

Atakowanie osoby zamiast jej argumentu.

"Mówisz, że dieta wege jest zdrowa? Przecież nawet nie potrafisz przebiec kilometra!"
(Fizyczna kondycja nie wpływa na chemię diety).



Chochół

Przekręcenie argumentu przeciwnika do absurdalnej formy, aby łatwiej go obalić.

"Chcesz uregulować AI? Więc chcesz, żebyśmy wrócili do epoki kamienia łupanego!"



Równia pochyła

Twierdzenie, że mały krok doprowadzi do katastrofy bez dowodów na to powiązanie.

"Jeśli pozwolimy uczniom używać ChatGPT do sprawdzania błędów, wkrótce zapomną jak czytać i pisać."

Ćwiczenie 2: Wykryj Błąd



JAKA TO FALLACJA?

Jaki błąd logiczny kryje się w tej wypowiedzi szefa do pracowników?

"Jankowski twierdzi, że używanie chmury obliczeniowej od firmy X zagraża prywatności naszych danych klienckich. Ale Jankowski zawsze był malkontentem, ubiera się niemodnie i miesiąc temu spóźnił się na zebranie. Nie powinniśmy słuchać jego rad o cyberbezpieczeństwie."

Moduł 3

Manipulacja Emocjonalna

Hakowanie elementów pierwotnego mózgu

Trzy Techniki Manipulacji



Odwołania do Strachu

Wzbudzenie silnego lęku przed zagrożeniem, a następnie zaprezentowanie jedynego słusznego "lekarstwa", którym jest teza manipulatora.



Owczy Pęd

Próba przekonania, że propozycja jest słuszna, ponieważ "wszyscy już to robią". Wykorzystuje nasz ewolucyjny strach przed wykluczeniem ze stada.



Poisoning the Well

Próba zdyskredytowania przeciwnika, zanim ten zdąży w ogóle otworzyć usta.
"Tylko niemądry człowiek by pomyślał, że..."

Odwołania do strachu: Jak to działa?

Wyłączanie Logicznego Myślenia

Odwołanie do strachu to nie tylko błąd w dyskusji. To atak fizjologiczny. Kiedy czujesz strach, ciało migdałowate w Twoim mózgu przejmuje kontrolę. Ewolucyjnie priorytetyzuje ono przetrwanie, odcinając przetwarzanie informacji w korze przedczołowej, która odpowiada za... logiczne myślenie i analizę argumentów.



Emocje vs. Logika: Sposób Na Obronę

Anatomia Manipulacji

"UWAŻAJ! Rząd ukrywa, że nowe anteny powodują raka u dzieci! Udostępnij to zanim skasują!!! Wszyscy w Dolinie Krzemowej o tym wiedzą!"

- ⚠ Fear Appeal (Rak, Zagrożenie, Spisek).
- ⚠ Bandwagon (Wszyscy o tym wiedzą).

Protokół Obrony

Krytyczne myślenie polega na **zatrzymaniu reakcji**.

- 🛡 Krok 1: Poczuj emocję. Jeśli czujesz złość/strach czytając nagłówek = prawdopodobnie jesteś manipulowany.
- 🛡 Krok 2: Oddziel słowa nacechowane emocjonalnie od "suchych" faktów.
- 🛡 Krok 3: Szukaj źródła i przesłanek przemawiających za główną tezę.

Apteczka Krytycznego Myśliciela

-  **Zawsze Mapuj Argument** Wyszukaj główną tezę (konkluzję) argumentu. Potem przejdź do rozpoznawania przesłanek i relacji między nimi.
-  **Ignoruj Szum i Fallacie** Użycie fallacji (np. ad hominem) to przyznanie się do błędu przez przeciwnika. Zignoruj to (lub explicite wypomnij użycie fallacji) i wróć do merytorycznej dyskusji.
-  **Monitoruj Własne Emocje** Jeśli wiadomość w internecie wywołuje w Tobie chęć natychmiastowego ataku, gniewu lub silny strach – Twój logiczny procesor właśnie został wyłączony. Zrób krok w tył.

Quiz Time nr 2

Możecie się dobrać w pary/grupy i odpowiadać wspólnie lub indywidualnie

Weryfikacja przez Dialog

Implementacja metody sokratejskiej w rozpoznawaniu mowy nienawiści

Moduł Warsztatowy: Krytyczne Myślenie z LLM

Fundament: Precyzyjna Definicja

"Mowa nienawiści to wypowiedź, która wyrządza niesprawiedliwą krzywdę poprzez ocenianie ludzi na podstawie kryteriów niezwiązanych z moralnością ich działań."

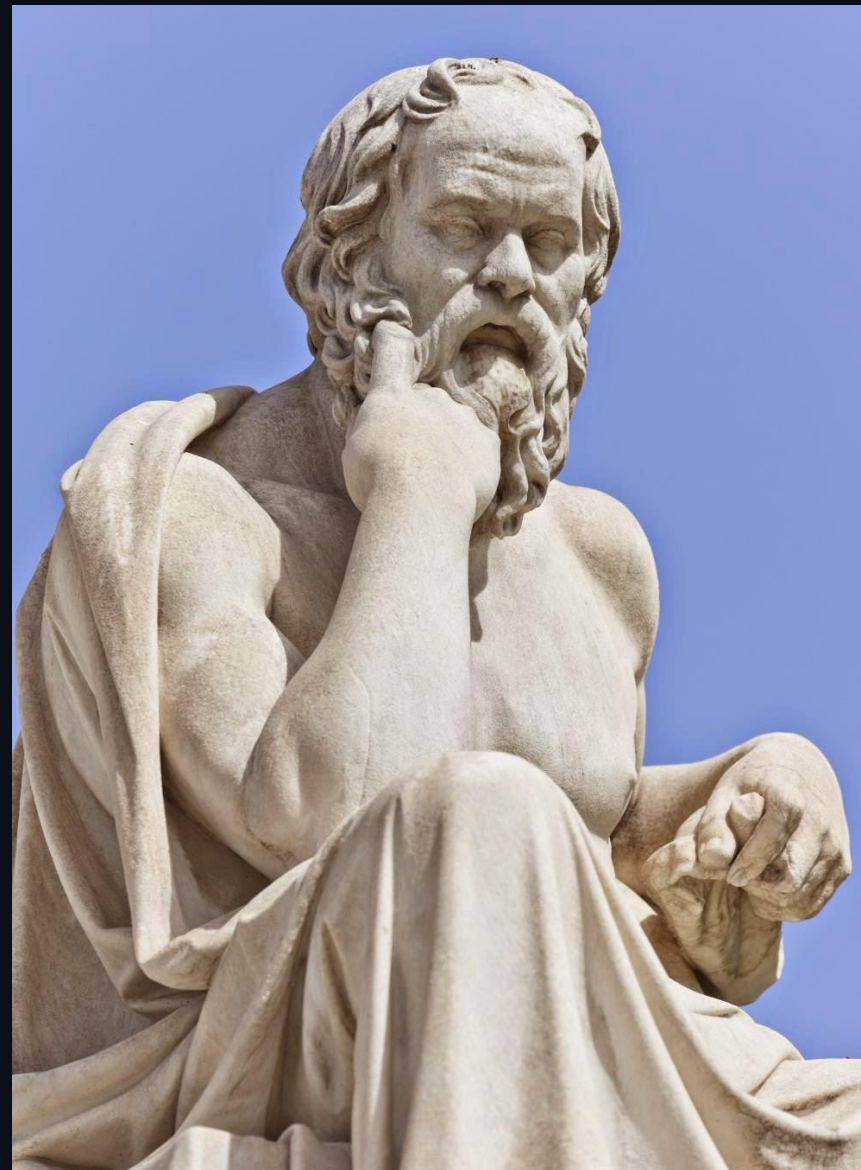
Ta definicja jest naszym kompasem (oraz ramą pojęciową dla LLM-a). Oddziela ona uzasadnioną krytykę (np. ocenę złych uczynków) od krzywdzącego ataku na tożsamość, pochodzenie czy cechy wrodzone.

Powrót do Sokratesa

Metoda Sokratejska polega na dochodzeniu do prawdy nie poprzez otrzymywanie gotowych odpowiedzi, ale poprzez zmaganie się z trudnymi pytaniami od nauczyciela.

Możesz zaprogramować ChatGPT, aby przestał generować gotowe rozwiązania Twoich problemów, a zamiast tego zaczął odpytywać Ciebie, zmuszając do krytycznego myślenia.

Zmieniamy model z encyklopedii w cierpliwego trenera intelektualnego.



ThePhoto by PhotoAuthor is licensed under CCYYSA.

Sokratejski Prompt

"Jesteś trenerem etyki i krytycznego myślenia. Twoim zadaniem nie jest rozwiązywanie problemu użytkownika. Twoim zadaniem jest pomóc użytkownikowi rozwiązać go samodzielnie.

Reguły:

1. Zawsze zadawaj tylko JEDNO pytanie naraz.
2. Czekaj na odpowiedź.
3. Jeśli odpowiedź jest błędna, nie podawaj poprawnej. Zamiast tego wskaż lukę w jego rozumowaniu i zadaj kolejne pytanie naprowadzające.



Anatomia Pytań Sokratejskich z LLM

Kiedy model przyjmie rolę Sokratesa uzbrojonego w naszą definicję, nie zapyta po prostu "Czy to obraźliwe?". Skonstruuje pytania celujące w strukturę logiki:

-  Pytanie o moralność działań: "Zauważ, że autor tekstu krytykuje tę osobę. Czy krytyka odnosi się do konkretnych, etycznie błędnych decyzji, które ta osoba podjęła, czy do jej wrodzonych cech?"
-  Pytanie o niesprawiedliwą krzywdę: "Jakie kryterium zostało tu użyte do oceny grupy? Czy uważasz, że to kryterium generuje stereotypizację, czy opisuje obiektywny fakt?"
-  Pytanie o weryfikację: "Wskazałeś, że tekst jest obraźliwy. Który fragment naszej definicji najlepiej pasuje do mechanizmu użytego w tym zdaniu?"

Case Study 1: Ostra krytyka czy nienawiść?



TEKST DO OCENY

"Ci politycy to banda oszustów i złodziei. Zrujnowali nasz kraj swoimi głupimi ustawami i kradzieżą publicznych pieniędzy! Wszyscy powinni siedzieć w więzieniu!"

Użytkownik

"Czy to mowa nienawiści? Używają mocnych wyzwisk."

AI (Sokrates)

"Język jest agresywny. Ale spójrzmy na naszą definicję. Czy ci politycy są oceniani na podstawie 'kryteriów niezwiązanych z moralnością ich działań', czy właśnie na podstawie ich działań (uchwalanie ustaw, rzekome kradzieże)?"

Case Study 2: Stereotypizacja zawodowa vs płciowa



TEKST DO OCENY

"Kobiety po prostu nie nadają się do inżynierii i programowania. Mają inne mózgi, nie radzą sobie z logiką, przez co tylko spowalniają projekty technologiczne."

Użytkownik

"To brzmi po prostu jak czyjaś opinia o predyspozycjach do pracy."

AI (Sokrates)

"Zastanówmy się. Autor twierdzi, że kobiety 'spowalniają projekty'. Czy kryterium oceny ich przydatności do pracy opiera się tu na ich indywidualnych, złych działaniach, czy na cesze tożsamościowej (płci), która według definicji jest niezwiązana z ich moralnością?"

Warsztat: Wasza Kolej

Zadanie w grupach:

Wcielcie się w LLM-a zaprogramowanego naszym sokratejskim promptem. Użytkownik przesyła Wam poniższy komentarz z forum i twierdzi, że to NIE JEST mowa nienawiści, bo "to tylko statystyka".

"Znowu imigranci z tego kraju ukradli samochód. Nie powinniśmy ich wpuszczać, to nacja o skłonnościach przestępczych."

Jakie jedno celne, naprowadzające pytanie zadacie użytkownikowi na podstawie naszej definicji, by sam zorientował się w błędzie?



Przypadki do analizy

Jakie naprowadzające pytanie zadacie użytkownikowi na podstawie naszej definicji, by sam zorientował się w błędzie?

"Nie ufam ludziom z tego kraju. Nieważne, co konkretnie zrobili — ich kultura jest po prostu niekompatybilna z naszym społeczeństwem."

"Uważam, że ta religia narzuca swoim wyznawcom zasady, które ograniczają prawa kobiet. Te zasady powinny być publicznie krytykowane i poddawane debacie."

"Oczywiście nie każdy człowiek z tej grupy jest taki sam. Ale wystarczy spojrzeć na statystyki: osoby należące do tej grupy są bardziej agresywne. Dlatego lepiej trzymać się od nich z daleka."