



www.aileaders-project.eu

ĆWICZENIE SCENARIUSZOWE – Jak opracować własnego własnego etycznego chatbota



Co-funded by
the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.

ĆWICZENIE SCENARIUSZOWE

- Jak opracować własnego własnego etycznego chatbota

01



Streszczenie 3

02



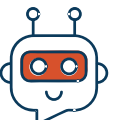
Wprowadzenie 3

03



Prezentacja narzędzi 4

04



Wytyczne dotyczące opracowania etycznego chatbota 5

05



Zajęcia praktyczne 7

06



Wnioski 8

07



Bibliografia 9

08



Materiały uzupełniające 9

• 01 Streszczenie



Rodzaj OER

Ćwiczenie scenariuszowe

Cel/przeznaczenie

Zapewnij praktyczny, niedrogi prototyp pokazujący, jak zbudować i wdrożyć **etycznego chatbota** w Google Colab.

Oczekiwane efekty kształcenia

Student będzie potrafił stworzyć funkcjonalnego chatbota przy użyciu narzędzi przyjaznych dla początkujących, z uwzględnieniem zasad etycznego projektowania.

Słowa kluczowe

- Etyczna sztuczna inteligencja
- Chatbot
- Prototyp
- Google Colab
- DialoGPT
- Hugging Face Transformers
- Gradio UI
- Moderacja treści
- Zastrzeżenia

Sugerowane Metodologiczne Podejście

Nauka oparta na problemach

• 02 Wprowadzenie



Generatywne chatboty przenoszą się z laboratoriów badawczych do codziennych punktów obsługi klienta, sal lekcyjnych i urządzeń osobistych, ale ich szybkie wdrażanie wyprzedza równie szybkie prace nad zabezpieczeniami **etycznej sztucznej inteligencji**.

Programiści, którzy chcą dostarczać użytecznych asystentów, stoją przed kluczowym problemem:

jak dodać nawet minimalną moderację treści i warstwy przejrzystości bez kosztownej infrastruktury lub zastrzeżonych interfejsów API?

W kolejnych sekcjach przedstawimy kilka wytycznych dotyczących wymagań, jakie powinien spełniać etyczny chatbot, a także praktyczne ćwiczenie pozwalające stworzyć własnego chatbota.

Ethical Chatbot (Demo)

This is a prototype chatbot with a very basic filter and simple disclaimers for demonstration. A production system should use a much more robust moderation pipeline.

user_input

Type your message...

output

Limpiar

Enviar

Marcar

• 03 Prezentacja narzędzi



Narzędzia wykorzystane w tym scenariuszu ćwiczeniowym

01 Python

Najczęściej używany język programowania w nauce o danych

02 Edytor języka Python

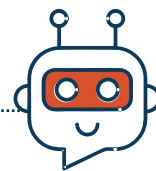
Google Colab lub Jupyter Notebook

03 Biblioteki i frameworki

- **Transformatory (Hugging Face):** ładowanie i uruchamianie wstępnie wytrenowanego modelu konwersacyjnego (DialoGPT-medium)
- **PyTorch (torch):** backend głębokiego uczenia się do wnioskowania modelowego
- **Gradio:** interfejs użytkownika bez kodowania do wdrażania i udostępniania chatbota

04 Zastosowane techniki

- Konfiguracja danych i kodu w bezpłatnym notebooku Colab z obsługą GPU
- Ładowanie wstępnie wyszkolonego modelu dialogowego i utrzymywanie stanu konwersacji
- Moderacja treści: czarna lista słów kluczowych (lub klasyfikator potokowy) w celu blokowania toksycznych wpisów
- Wykrywanie tematów wrażliwych: skanowanie słów kluczowych → automatyczne zastrzeżenia („skonsultuj się z profesjonalistą”)
- Generowanie odpowiedzi chatbota: pobieranie próbek z DialoGPT z kontekstem historii
- Szybkie wdrażanie: interfejs Gradio z pojedynczą komórką, który generuje publiczny link do testowania



• 04 Wytyczne dotyczące opracowania etycznego chatbota

Wymagania dotyczące etycznego chatbota

01 OKREŚL CEL I ZAKRES DZIAŁANIA CHATBOTA

Określ przypadek użycia:

- Określ, co powinien robić Twój chatbot (np. odpowiadać na pytania dotyczące obsługi klienta, ułatwiać wsparcie w zakresie zdrowia psychicznego lub udzielać ogólnych odpowiedzi na pytania).

Wyjaśnij ograniczenia:

- Wyraźnie określ, czego chatbot nie może robić (np. udzielać porad medycznych, obsługiwać transakcji finansowych).

02 USTAL WYTYCZNE ETYCZNE I ZACHOWANIOWE

Opracuj „Kartę etyczną” lub politykę:

- Określ zasady, których powinien przestrzegać Twój chatbot (np. unikanie szkodliwych treści, poszanowanie prywatności użytkowników).
- Uwzględnij uczciwość, bezpieczeństwo, przejrzystość i poszanowanie różnorodności.

Stwórz „Kodeks postępowania” dla swojego chatbota:

- Jak powinien reagować na treści zawierające nienawiść lub molestowanie?
- Jakie stanowisko zajmuje w sprawie dezinformacji lub fałszywych informacji?
- Jak powinien postępować w przypadku prób o podanie danych osobowych lub wrażliwych?

Skonfiguruj mechanizmy awaryjne:

w sytuacjach, gdy chatbot otrzymuje nielegalne lub szkodliwe polecenia, zaprojektuj go tak, aby odpowiadał odmową lub bezpieczną alternatywą (np. kierując użytkownika do wykwalifikowanych specjalistów).

03 SZKOLENIE MODELU

Źródło etycznych danych szkoleniowych:

- Upewnij się, że dane nie naruszają prywatności ani praw autorskich.
- Staraj się uzyskać reprezentatywne dane, które obejmują różne języki, regiony i kultury, aby zminimalizować stronniczość.
- Filtrowanie i wstępne przetwarzanie:
- Usuń lub oznacz treści potencjalnie szkodliwe lub stronnicze.
- Rozważ użycie gotowych narzędzi lub wewnętrznych filtrów do identyfikacji mowy nienawiści, treści o charakterze jednoznacznym lub danych osobowych.

Wybór modelu:

- Zrównoważ złożoność modelu z interpretowalnością i wykorzystaniem zasobów.
- Uczenie się ze wzmocnieniem na podstawie informacji zwrotnych od ludzi (RLHF) lub podobne podejścia mogą pomóc w udoskonaleniu zachowania modelu po szkoleniu.

04 ZAPEWNIJ UŻYTKOWNIKOM PRZEJRZYSTOŚĆ

Ujawniaj charakter swojego chatbota:

- Wyraźnie poinformuj użytkowników, że mają do czynienia z systemem opartym na sztucznej inteligencji.
- Przedstaw zastrzeżenia dotyczące wszelkich ograniczeń i możliwych niedokładności.

Wyjaśnij wykorzystanie danych:

- Poinformuj użytkowników, w jaki sposób ich dane są przechowywane, przetwarzane i wykorzystywane do ulepszania systemu.
- Przestrzegaj przepisów dotyczących ochrony danych (np. RODO, CCPA) i zapewnij łatwy dostęp do polityki prywatności.

Zapewnij łatwe „wyjście”:

- Poinformuj użytkowników, że w dowolnym momencie mogą wstrzymać lub zrezygnować z gromadzenia danych.
- Zapewnij mechanizm umożliwiający usunięcie lub anonimizację danych na żądanie.

05 POSTĘPUJ W PRZYPADKACH GRANICZNYCH I WRAŻLIWYCH TEMATACH

Odmowa lub bezpieczne zakończenie:

- W przypadku nielegalnych, szkodliwych lub bardzo wrażliwych próśb (np. porad medycznych, prawnych, finansowych) chatbot powinien:
 - Przedstawić zastrzeżenie i częściowe informacje wraz z odniesieniami lub
 - Skierować użytkownika do profesjonalnej pomocy.

Kontekst zdrowia psychicznego lub samookaleczeń:

- Przygotuj bezpieczne odpowiedzi: okaż empatię, zachęć użytkownika do skorzystania z profesjonalnej pomocy i podaj dane kontaktowe służb wsparcia (jeśli to możliwe).

Sytuacje awaryjne:

- Wyraźnie zaznacz, że chatbot nie jest przystosowany do obsługi sytuacji awaryjnych.
- Podaj instrukcje lub dane kontaktowe służb ratowniczych, jeśli pojawią się rozpoznawalne sygnały zagrożenia.

06 ZACHOWAJ ZGODNOŚĆ Z WYMOGAMI PRAWNYMI I ETYCZNYMI

Zgodność z przepisami:

- Bądź na bieżąco z przepisami dotyczącymi sztucznej inteligencji, prywatności danych i usług cyfrowych w regionach, w których będzie działał Twój chatbot.
- Wdroż solidne mechanizmy ochrony danych i uzyskiwania zgody.

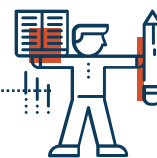
Kwestie odpowiedzialności:

- Jasno określ, kto ponosi odpowiedzialność, gdy bot dostarcza nieprawidłowe lub szkodliwe informacje.
- Opublikuj zastrzeżenia szczegółowo opisujące zakres działania chatbota (np. „nie jest licencjonowanym pracownikiem służby zdrowia”).

Zachowaj jasną wizję etyczną

- Utrzymanie etycznego chatbota to ciągły proces. Okresowo weryfikuj swoje zasady i upewnij się, że są one zgodne z nowymi przepisami, wartościami społecznymi i opiniami społeczności. Spójność i przejrzystość aktualizacji pomagają budować zaufanie. Śledź wskaźniki wydajności i uwzględniaj opinie użytkowników, aby osiągnąć ciągłość.

• 05 Ćwiczenia praktyczne



Opis rozwiązania (postępuj zgodnie z instrukcjami zawartymi w dostarczonym notatniku)

To ćwiczenie pokaże Ci, jak stworzyć prototyp chatbota przy użyciu wyłącznie darmowych narzędzi: notatnika Google Colab, który łączy model Hugging Face Transformers (np. DialoGPT), dodaje do niego proste zasady moderacji treści i zastrzeżenia dotyczące wrażliwych tematów, a następnie wyświetla wynik za pomocą interfejsu użytkownika Gradio bez konieczności pisania kodu. W niecałą godzinę zobaczysz, jak przejść od koncepcji do linku, który można udostępnić,

udowadniając, że szybkie wdrożenie i odpowiedzialny projekt nie muszą się wzajemnie wykluczać. Udostępniamy gotowy do użycia **notatnik Google Colab**, który krok po kroku przeprowadzi Cię przez proces wdrażania etycznego chatbota. Notatnik ma przejrzystą strukturę podzieloną na sekcje, dzięki czemu można zrozumieć i modyfikować każdą warstwę rozwiązania.

▶ Wezwanie do działania

Otwórz notatnik *Ethical Chatbot Demo*, wykonaj kolejno każdą sekcję, a w ciągu kilku minut będziesz mieć działającego chatbota z podstawowymi zabezpieczeniami etycznymi, gotowego do zaprezentowania kolegom z zespołu lub powtórzenia w celu wykorzystania w produkcji.

• 06 Wnioski



Kluczowe spostrzeżenia

- 01** Jasne określenie, co chatbot powinien, a czego nie powinien robić, jest najskuteczniejszym zabezpieczeniem przed nadużyciami w przyszłości.
- 02** Stronniczość danych szkoleniowych przekłada się bezpośrednio na zachowanie modelu; różnorodne zbiory danych, które szanują prywatność, są niepodważalne.
- 03** Filtr moderacji w czasie rzeczywistym wychwytuje różne rodzaje błędów.
- 04** Przezroczystość jest koniecznością. Użytkownicy, którzy wiedzą, że rozmawiają z AI, rozumieją ograniczenia, kontrolują swoje dane i są bardziej wyrozumiali w przypadku sporadycznych błędów.
- 05** Zarządzanie musi być ciągłe, aby zapewnić nie tylko dobrą wydajność, ale także spełnienie wymogów prawnych i etycznych.

Kluczowe wnioski

- 01** Łatwiej jest złagodzić zbyt surowe zabezpieczenia niż wprowadzać moderację po tym, jak szkodliwe treści dotarły do użytkowników.
- 02** Śledź wskaźniki szkodliwych wyników, sprawiedliwość demograficzną i satysfakcję użytkowników, a następnie powiąż aktualizacje modelu z konkretnymi celami.
- 03** Mniejsze, interpretowalne modele z jasnymi zabezpieczeniami często przewyższają większe modele typu „czarna skrzynka” pod względem bezpieczeństwa i zgodności z przepisami.
- 04** Narzędzia i zasady nie sprawdzają się bez zespołu, który przedkłada dobro użytkowników nad szybkość lub nowatorskość.
- 05** Włącz etykę do każdego sprintu.

• 07 Bibliografia



- European Commission. (2024). EU Artificial Intelligence Act: <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>
- Hugging Face Transformers Documentation: <https://huggingface.co/docs/transformers/en/index>
- Gradio documentation: <https://www.gradio.app/docs>



• 08 Materiały uzupełniające



Plik Jupyter Notebook (IPYNB)

Dołączono szczegółowy notatnik Python z komentowanym kodem, który przeprowadza użytkownika przez każdy etap procesu.

Notebook zawiera

Konfigurację środowiska	Instalacja transformatorów, torch i gradio; sprawdzenie GPU.
Ładowanie modelu	Pobieranie otwartego modelu konwersacyjnego DialoGPT-medium (lub dowolnego innego modelu) z Hugging Face i zainicjowanie bufora konwersacji.
Podstawowy filtr moderacyjny	Wdraża prostą czarną listę słów kluczowych i pokazuje, jak zamienić ją na potok toksyczności Hugging Face w celu uzyskania silniejszego filtrowania.
Ograniczenia dotyczące tematów wrażliwych	Skanuje dane wprowadzone przez użytkownika pod kątem słów kluczowych związanych z medycyną, prawem lub finansami i automatycznie dodaje zastrzeżenie „Skonsultuj się z profesjonalistą”.
Logika generowania odpowiedzi	Koduje polecenie użytkownika, dołącza historię czatu i pobiera próbki z modelu, aby wygenerować odpowiedź uwzględniającą kontekst.
Wdrożenie interfejsu użytkownika Gradio	Zawiera funkcję odpowiedzi w interfejsie internetowym bez kodu; uruchomienie komórki zwraca publiczny link, który można natychmiast udostępnić.
Wskazówki dotyczące dostosowywania	Komentarze wbudowane pokazują, gdzie można podłączyć inny model, rozszerzyć listę filtrów lub dodać analizy.



leaders

Śledź naszą podróż



www.aileaders-project.eu



Co-funded by
the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.