
Kraków
ZADANIE PUBLICZNE WSPÓŁFINANSOWANE ZE ŚRODKÓW MIASTA KRAKOWA
FUNDACJA OPTIMUM PARETO
AI w pracy: modele językowe
broszura informacyjna

Spis treści

PRZEDSŁOWIE

Celem niniejszej broszury informacyjnej jest upowszechnienie wiedzy oraz dobrych praktyk stosowania modeli językowych w środowiskach profesjonalnych. Zostaje ona opublikowana dokładnie rok od premiery aplikacji Chat GPT, która zrewolucjonizowała dostęp do sztucznej inteligencji. Korzystanie z niej nie wymaga już specjalistycznej wiedzy, ponieważ zwykły dialog pełni funkcję interfejsu, a technologia ta dobrze radzi sobie z wyręczaniem nas w wielu codziennych zadaniach. Jednak powszechna adaptacja tej technologii nie jest równoznaczna ze świadomym korzystaniem z niej, co jest bardzo ważne w szczególności w środowisku pracy.

Transparentność działania modeli językowych, nawet względem jej twórców, pozostawia dużo do życzenia i wciąż jest obiektem badań. Technologia ta cały czas też ewoluuje, co utrudnia wypracowanie stabilnych praktyk skutecznego stosowania jej w miejscu pracy.

W odpowiedzi na te wyzwania postanowiliśmy zaprosić do dyskusji osoby reprezentujące środowiska biznesowe, akademickie i prawne, aby zebrać różnorodne perspektywy na ten sam fenomen i za pomocą inteligencji zbiorowej opracować rzetelne, dobre praktyki, które można bez zastrzeżeń rozpowszechniać wśród przedsiębiorców, dyrektorów, pracowników i osób zainteresowanych świadomym korzystaniem z dialogicznej sztucznej inteligencji. Spotkania tego rodzaju przeprowadziliśmy wcześniej w październiku 2023 w Warszawie, a następnie rozwinęliśmy je w niniejszej edycji (listopad 2023), także w Krakowie. Kluczowe wnioski ze spotkań zweryfikowane przez specjalistów Fundacji Optimum Pareto zawarte są w niniejszej broszurze.

Ciekawej lektury w imieniu Fundacji Optimum Pareto życzy autor broszury i prezes zarządu - Marcin Woźniak

FUNDACJA
OPTIMUM
PARETO

SŁOWNICZEK POJĘĆ

Model językowy

Model językowy to program komputerowy lub algorytm oparty na sztucznej inteligencji, który ma zdolność przewidywania i generowania tekstu w języku naturalnym. Trenowane na dużej ilości danych tekstowych, modele językowe uczą się statystyk i struktury języka, umożliwiając im tworzenie sensownych i gramatycznie poprawnych zdań. Model językowy analizuje sekwencje słów (lub tokenów) w kontekście, ucząc się prawdopodobieństw występowania kolejnych słów na podstawie poprzednich. Dzięki temu może generować tekst na podstawie danego początkowego fragmentu lub kontekstu.

Prompt

"Prompt" to słowo lub zdanie, które jest używane do inicjowania lub wyzwalania odpowiedzi od systemu opartego na generatywnej sztucznej inteligencji. Jest to sposób, w jaki użytkownik komunikuje się z modelem, zwykle prosząc o odpowiedź na pytanie, stworzenie tekstu, opisanie obrazu lub wykonanie innej czynności. Treść promptu ma istotny wpływ na to, jakie informacje zostaną zawarte w odpowiedzi modelu. Wprowadzenie różnych promptów może skutkować zróżnicowanymi i nieprzewidywalnymi odpowiedziami od modelu.

SŁOWNICZEK POJĘĆ

Token

Token w kontekście modeli językowych to podstawowa jednostka w przetwarzaniu języka naturalnego. Może on reprezentować pojedynczy znak, słowo lub inną jednostkę tekstową. Różne modele mają różną tokenizację. Modele językowe analizują sekwencje tych tokenów w kontekście, przewidując prawdopodobieństwo występowania kolejnych tokenów na podstawie poprzednich. Dzięki temu mogą generować tekst na podstawie danego początkowego fragmentu lub kontekstu. Wykorzystywana liczba tokenów ma wpływ na koszty obliczeniowe i limity ilości treści, które można przekazać do modelu podczas jednego zapytania

Okno kontekstowe

Okno kontekstowe to fragment informacji, które są widoczne dla modelu w danym momencie. Można o tym myśleć jako o pamięci roboczej modelu. Zdarza się, że podczas rozmowy z nami model zapomina o tym co wcześniej napisaliśmy. Może to wynikać z tego, że informacje te wypadły z okna kontekstowego modelu. Wyzwaniem dla skutecznego stosowania modelu jest zawarcie wystarczająco istotnych informacji w taki sposób nie przekraczać limitu tokenów, które model może obsłużyć w jednym zapytaniu. Techniki wspomagające optymalne zarządzanie oknem kontekstowym są kluczowe dla uzyskania precyzyjnych i adekwatnych odpowiedzi od modelu generatywnego.

SŁOWNICZEK POJĘĆ

Halucynacje

Halucynacjami określa się generowanie nieprawdziwych informacji przez sztuczną inteligencję. Informacje te mogą brzmieć przekonująco dla człowieka np. posiadać prawdopodobnie brzmiące nazwy, lub przypisanie autorstwa cytatu osobie, która zajmowała się podobną tematyką, ale nigdy nie napisała tego co twierdzi model językowy. Nie wynaleziono jeszcze algorytmu, który skutecznie odróżniałby prawdę od konfabulacji. Modele językowe trenowane są na zadaniu jakim jest przewidzenie jaki kolejny token w ciągu spowoduje akceptację zdania przez człowieka. W związku z tym modele dziedziczą także pewne błędy poznawcze, charakterystyczne dla naszego myślenia. Konieczne jest zatem weryfikowanie treści produkowanych przez model.

Multimodalność

Multimodalność modelu to połączenie różnych form przetwarzania informacji. Mogą to być dane takie jak dźwięk, tekst, obraz wideo, dane liczbowe, kod komputerowy, dane pochodzące z czujników, aplikacji i inne rodzaje danych. Multimodalność pozwala na analizę większej ilości informacji o świecie rzeczywistym, a także na generowanie odpowiedzi, które łączą w sobie różne media. Wiele osób upatruje kolejne przełomy w sztucznej inteligencji właśnie w aspekcie multimodalności. Modele językowe dobrze radzą sobie z tłumaczeniem między językami, ale nie musi się to ograniczać do tego, co my jako ludzie traktujemy jako język, co może rodzić kolejne problemy związane z wyjaśnialnością działania multimodalnych systemów.

POPULARNE MODELE JĘZYKOWE

Chat GPT (OpenAI) - platforma z wtyczkami oferująca multimodalną analizę oraz generację tekstu i obrazu

Bing (Microsoft) - model z wyszukiwarką oparty na podobnych technologiach jak Chat GPT

Bard (Google) - multimodalny chat oparty na modelu językowym PaLM 2

Claude (Anthropic): model językowy znany z dużego okna kontekstowego

Elicit (Ought) - dedykowany model do przeglądania literatury naukowej

Perplexity (Perplexity AI) - wyszukiwarka zintegrowana z różnymi modelami językowymi

Llama 2 - model open source firmy Meta

Trurl (Voice Lab) - polski, multimodalny model językowy

INNE PRZYDATNE NARZĘDZIA

Litmaps - aplikacja, która pozwala na wyszukiwanie i wizualizację powiązań artykułów naukowych poprzez graf cytowań

Lakera - program, który zabezpiecza modele językowe przed ryzykami technicznymi takimi jak wstrzyknięcie promptu, utrata danych, czy niebezpieczne treści

Hackstery - blog o cyberbezpieczeństwie z newsletterem na temat zagrożeń związanych z AI

HeyGen - aplikacja do tworzenia awatarów, animacji, konwertowania tekstu na mowę, synchronizacji dźwięku z ruchem warg

ElevenLabs - model pozwalający na przekształcanie tekstu na mowę w różnych językach oraz klonowanie głosu

Whisper - to program do zamiany mowy na tekst służący np. do robienia transkrypcji i napisów do filmów

Swarmcheck - program, który w oparciu o interaktywne mapy argumentacji wspiera dyskusje, decyzje, tworzenie wiarygodnych szacunków liczbowych i priorytetyzację. Program Swarmcheck wspierał interdyscyplinarne dyskusje prowadzące do stworzenia niniejszej broszury

HuggingFace - to platforma oferująca biblioteki i narzędzia do tworzenia, trenowania i wdrażania modeli językowych opartych na sztucznej inteligencji

DeepL - program do tłumaczenia automatycznego tekstu

Scikit-learn - to biblioteka open source uczenia maszynowego dla języka Python. Zawiera ona różne algorytmy klasyfikacji, regresji i klastrowania

LlamaIndex - framework do łączenia własnych źródeł danych z dużymi modelami językowymi

LangChain - framework upraszczający tworzenie aplikacji wykorzystujących duże modele językowe

SKUTECZNE ZASTOSOWANIA MODELI JĘZYKOWYCH

Edycja tekstu - dzięki modelom językowym możliwe jest wykonywanie wielu zadań polegających na przetwarzaniu języka naturalnego. Do często wykonywanych zadań należy ***tłumaczenie*** między różnymi językami, a także dialektami, ***copywriting*** np. treści ogłoszeń i postów w mediach społecznościowych, ***przepisywanie tekstu w zadanym stylu*** od raportów technicznych i sprawozdań do tekstów literackich jak tworzenie wierszy czy piosenek, *a także* ***redagowanie i wysyłka wiadomości email***. Modele językowe mogą także poprawić zrozumiałość tekstu przy zachowaniu najważniejszych informacji. Dzięki temu pomocne są przy szeroko pojętych procesach komunikacji w tym działaniach marketingowych, przygotowaniu scenariusza wystąpienia lub treści prezentacji.

Podsumowanie - Wiele zadań wymaga syntezy długiego tekstu. W przeprowadzonym badaniu streszczeń abstraktów zlecanych na platformie Amazon Mechanical Turk wykazano, że 33-46% pracowników korzystało z modeli językowych przy wykonywaniu tego zadania.

SKUTECZNE ZASTOSOWANIA MODELI JĘZYKOWYCH

Pomoc w nauce - ze względu na rozległą wiedzę jaka reprezentowana jest w modelach językowych, mogą one wspomagać proces szkolenia pracowników, oraz tworzyć dedykowane materiały edukacyjne. Do popularnych zastosowań należy nauka języka obcego, nauka programowania, metoda dialogu Sokratejskiego w analizie przekonań.

Interfejs do innych programów - wielu użytkownikom znacznie łatwiej może przyjść powiedzenie tego co chcą osiągnąć w danym programie komputerowym, zamiast uczyć się dedykowanego interfejsu. Dlatego zastosowanie modeli językowych do generowania i edycji grafik, tworzenia dokumentów lub prezentacji. Wraz z szerszą adaptacją tej technologii, możemy się spodziewać większej integracji tego rodzaju systemów z innymi programami oraz większego przyzwyczajenia użytkowników do takiej formy interakcji.

Wsparcie w podejmowaniu decyzji - o ile pełna automatyzacja decyzji i procesów wciąż może być jeszcze ryzykowna, bez odpowiedniej implementacji, to wsparcie decyzji na etapie brainstormingu, szukania alternatyw oraz przedstawiania konstruktywnej krytyki, jak najbardziej leży w zasięgu kompetencji współczesnych modeli językowych. O ile ostateczne decyzje powinien podejmować człowiek, który jest za nie odpowiedzialny, to dostarczenie większej ilości perspektyw i dodatkowej analizy może być cennym zasobem.

Wsparcie prac programistycznych - wiele współczesnych modeli językowych uczonych było także na danych zawierających otwarte oprogramowanie oraz kod. O ile ze złożonymi poleceniami nadal nie radzą sobie świetnie, to dla prostych zadań jak sprawdzanie błędów, tworzenie dokumentacji sprawdza się dobrze. Przy stosowaniu warto sprawdzić, czy model z jakiego korzystamy trenowany był na aktualnych bibliotekach programistycznych, lub ma do nich dostęp.

SKUTECZNE ZASTOSOWANIA MODELI JĘZYKOWYCH

Asystent klienta - modele językowe mogą odpowiadać na typowe pytania klienta, dostarczać przydatnych instrukcji oraz samouczków udzielać wsparcia technicznego oraz wspierać proces reklamacji, jednak wymaga to odpowiedniej konfiguracji np. douczenia modelu na danych firmy, zastosowania fine tuningu, techniki RAG (Retrieval Augmented Generation) lub odpowiednich reguł odpowiadania na pytania. W tym zastosowaniu należy zachować szczególną ostrożność, czy model nie halucynuje. Dobrym rozwiązaniem może być też ustawienie niskiej "temperatury" czyli parametru odpowiadającego za kreatywność typowej odpowiedzi.

Przeszukiwanie i klasyfikacja danych - wklejanie tekstu, w którym znajduje się dana odpowiedź pozwala na odpytywanie modelu o informacje znajdujące się w tym tekście oraz tworzenie dobrze opisanych zbiorów danych. Jest to możliwe pod warunkiem, że tekst mieści się w oknie kontekstowym modelu. Przykładem takiej techniki jest In-Context Retrieval-Augmented Language Models.

Przegląd literatury - narzędzia takie jak Elicit, wedle twórców mają około 90% poprawności w odpowiadaniu na pytania przekrojowe związane z jakąś dziedziną naukową. Nie jest to poziom, który przewyższałby opracowanie eksperta dziedzinowego, jednak może podsunąć nam pewien ogląd przedmiotu, co może być przydatne dla interdyscyplinarnego spojrzenia na dane zagadnienie. Wyszukiwarka Perplexity także ma dedykowany tryb do literatury naukowej. Należy przy tym kierować się zasadą ograniczonego zaufania, narzędzia takie jak Chat GPT mogą np. halucynować bibliografię lub powielać pewne popularne błędy. Warto posiłkować się takimi narzędziami jak Litmaps, który pozwala na wyszukiwanie artykułów powiązanych ze znanym nam artykułem przez siatkę cytowań.

POTENCJALNY WPŁYW MODELI JĘZYKOWYCH NA RYNEK PRACY

Upowszechnienie się modeli językowych niewątpliwie wpłynie na rynek pracy. Analiza wpływu wielkich modeli językowych na rynek pracy w USA *GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models* sugeruje, że **około 47-56% wszystkich zadań wykonywanych przez pracowników mogłaby zostać wykonana znacznie szybciej przy tym samym poziomie jakości.** W rezultacie, niektóre stanowiska mogą ulec redukcji, ponieważ mniej osób jest w stanie wykonywać podobną pracę. Przy tym całkowite zastąpienie zawodów, wydaje się być odleglejszą perspektywą. Biorąc pod uwagę zwiększenie wydajności pracy jaka związana jest umiejętnością skutecznego korzystania z modeli językowych, stawać się to będzie **jedną z kluczowych kompetencji wpływających na atrakcyjność pracownika na rynku pracy**. Ze względu na ilość potencjalnych zastosowań modeli językowych, **inżynieria promptów ma większą szansę stać się umiejętnością wykorzystywaną na wielu stanowiskach**, jak np. obsługa arkusza kalkulacyjnego, niż samodzielnym zawodem.

Sztuczna inteligencja może być także narzędziem, które ułatwiać będzie zakładanie własnej działalności gospodarczej i startupów. Ponieważ daje to szansę na zmniejszenie kosztów prowadzenia przedsiębiorstwa przez **mniejsze zespoły wspierane przez AI.** Modele językowe mogą pomóc tworzyć strategie biznesowe, generować treści marketingowe, a nawet tworzyć i prototypować proste aplikacje.

Także większe firmy entuzjastycznie nastawione są na korzyści płynące ze stosowania AI. Niemal wszystkie osoby pracujące na stanowiskach menadżerskich (96%) uznają, że **generatywna sztuczna inteligencja jest kluczowym tematem rozmów w czasie posiedzeń zarządów**. Według raportu Capgemini aż **40% organizacji z różnych branż stworzyło już zespoły i budżet dedykowane tej technologii, a kolejne 49% rozważa zrobienie tego w ciągu następnych 12 miesięcy.**

METODY USPRAWNIANIA MODELI JĘZYKOWYCH

Inżynieria promptów - to ogólna nazwa na techniki poprawiania wyników generowanych przez sztuczną inteligencję, za pomocą wprowadzania skutecznych danych wejściowych np. odpowiednio sformułowanych poleceń. Najlepiej działające prompty odkrywane są przez badaczy modeli językowych oraz społeczność użytkowników, którzy za pomocą metody systematycznego podejmowania prób i błędów odkrywają skuteczność różnych technik na podstawie standaryzowanych testów. Jednak nie zawsze cel jaki chcemy osiągnąć posiada taki wykonany test, dlatego pozostaje nam często eksperymentować na własną rękę i porównywać wyniki własne oraz wymieniać się spostrzeżeniami z innymi. W internecie można znaleźć wiele **bezpłatnych poradników** jak zacząć swoją przygodę z inżynierią promptów. Możemy też zapytać model o to jak napisać dobry prompt lub go poprawić, aby skutecznie osiągnąć nasz cel. Poniżej, poza kilkoma metodami wskazanymi na końcu, zostaną przedstawione techniki głównie oparte o inżynierię promptów wraz z ich wyjaśnieniem.

METODY USPRAWNIANIA MODELI JĘZYKOWYCH

Ogólne wskazówki - Warto zacząć od jak najprostszych promptów, a dopiero później próbować je modyfikować. Prompt powinien zawierać wyraźne polecenie. Dobrze jest czytelnie odseparować instrukcje od treści, którą podajemy do analizy, np. za pomocą symboli ###. Im Polecenia powinny być konkretne i precyzyjne. Sformułowanie promptu jako nakaz jest skuteczniejsze niż zakazy. Np. zamiast biernego “nie rób X” lepiej zadziała aktywne “unikaj X”.

Zadawanie pytań po angielsku - jest techniką często stosowaną przez osoby, które przyszły na spotkania dyskusyjne. Większa skuteczność modelu w języku angielskim wynika prawdopodobnie z faktu, że duża ilość danych treningowych jest właśnie w języku angielskim. Dodatkowo, większość badań na temat skuteczności sformułowań różnego rodzaju promptów jest prowadzona w języku angielskim. Na koniec takiego promptu można dopisać, aby model odpowiedział po polsku, lub otrzymany tekst w języku angielskim można przetłumaczyć na polski w następnym kroku.

Few shot prompting - co można luźno przetłumaczyć jako podanie kilku przykładów w zapytaniu. W odróżnieniu od zadawania naszego pytania wprost, podajemy kilka przykładów, które podpowiadają modelowi jak wygląda oczekiwana odpowiedź. Dzięki przykładom model może zarówno skupić się na odpowiednim temacie analizy jak i dostosować się do oczekiwanej formy odpowiedzi. Przykłady negatywne, które jako takie są oznaczone i podane wśród przykładów pozytywnych (także oznaczonych) są skuteczne.

Wnioskowanie “krok po kroku” - Dla niektórych działań, gdzie oczekujemy poprawnego wyciągania wniosków, komenda “Zastanów się krok po kroku.” dodana do promptu, sprawia, że model językowy w odpowiedzi opisuje proces dochodzenia do konkluzji. Dzięki temu model ma mniejszą szansę na błąd. Badania nad automatycznym generowaniem promptów, znalazły najskuteczniejszą angielską wersja promptu: “Let's work this out in a step by step way to be sure we have the right answer”.

METODY USPRAWNIANIA MODELI JĘZYKOWYCH

Panel ekspertów - popularną metodą uzyskiwania odpowiedzi jest rozpoczęcie interakcji od wskazania “jesteś ekspertem z dziedziny X”. Taką instrukcję można także wykorzystać wraz z myśleniem krok po kroku, aby stworzyć dialog lub panel ekspertów debatujący nad danym zagadnieniem. Eksperci mogą reprezentować różne dziedziny, ale także role w dialogu np. generowanie pomysłów, weryfikacja błędów, proponowanie alternatyw lub decydowanie. Pomocne jest określenie ilości rund dyskusji. Przykładowo:

Trzej eksperci dyskutują nad pytaniem zapisując swoje analizy krok po kroku. Pierwszy ekspert proponuje rozwiązania, drugi ekspert odnosi się do pomysłów dostarczając konstruktywnej krytyki, trzeci ekspert rozwija najlepsze propozycje, aby dostarczyć dobrej odpowiedzi. Eksperci odnoszą się wzajemnie do tego co napisali inni wskazując dobre i złe argumenty. Dyskusja nad pytaniem trwa 3 rundy, a w 4 rundzie podejmowana jest decyzja. Zapisz dyskusję w formie tabeli. Pytanie to…

Custom instructions - coraz więcej interfejsów modeli językowych jak np. Chat GPT Plus lub Perplexity oferuje możliwość personalizacji swojej interakcji. Polega to na stworzeniu opisu siebie jako użytkownika aplikacji oraz wskazanie w jakim celu najczęściej się z niej korzysta.

Fine tuning - polega na dostosowywaniu parametrów już wytrenowanego modelu uczenia maszynowego do nowego zadania lub zbioru danych, aby przyspieszyć uczenie i osiągnąć lepszą wydajność na nowym zadaniu.

Retrieval augmented generation (RAG) - to technika, która wykorzystać dane na których model nie był wcześniej trenowany (np. dane firmy) do interakcji z modelem. Zapytanie użytkownika najpierw uruchamia proces wyszukiwania informacji w dostarczonej bazie danych, a następnie łączy te dane z pierwotnym promptem, bez potrzeby zmiany parametrów modelu. Dzięki temu jest to tańsza technika niż fine tuning. Jej skuteczność zależy od dobrej indeksacji i opisu naszych danych (w czym także może pomóc model językowy).

RYZYKA ZWIĄZANE ZE STOSOWANIEM MODELI JĘZYKOWYCH W PRACY

Sztuczna inteligencja to jeszcze młoda technologia, która już przekracza zdolności ludzkie w wielu aspektach.
Wraz z rozwojem jej kompetencji pojawiać się będą także nowe ryzyka, w tym o charakterze globalnym. Wymaga to merytorycznej i demokratycznej debaty. Jednak już teraz niektóre ryzyka mogą mieć wpływ na naszą pracę. Musimy nauczyć się krytycznie patrzeć na produkowane treści i umiejętnie weryfikować twierdzenia generowane przez AI.

- **Wprowadzanie w błąd** polega na generowaniu informacji przez model, które mają formę wiarygodnych informacji, lecz nie mają oparcia w rzeczywistości. Może to być wynikiem halucynacji modelu lub celowego tworzenia fake-news i innych rodzajów dezinformacji lub propagandy przez innych.

- **Nadmierne poleganie na autorytecie AI,** wiąże się z wpływem jaki mogą mieć systemy AI na decyzje ludzi. Eksperymenty prowadzone przez psychologów pokazują, że ludzie często bezkrytycznie polegają na poleceniach wydawanych przez sztuczną inteligencję.

RYZYKA ZWIĄZANE ZE STOSOWANIEM MODELI JĘZYKOWYCH W PRACY

- **Uzależnienie i zabieranie uwagi**, może występować w wyniku konkurencji między usługodawcami o czas i uwagę klientów. Modele językowe mają potencjał do uczenia się preferencji użytkowników i wykorzystywania tej wiedzy do uzależnienia użytkowników od interakcji z modelem i zabierania uwagi od innych, ważniejszych zadań. Podobnie jak w przypadku uzależnienia od mediów społecznościowych, nadmierna interakcja z modelami językowymi niesie ze sobą ryzyko niebezpiecznego wpływu na układu nagrody w ludzkim mózgu. Podawanie poszczególnym osobom takich informacji jakie chcą usłyszeć, a które niekoniecznie są prawdziwe, może prowadzić do pogłębienia baniek informacyjnych oraz większej izolacji społecznej i innych negatywnych konsekwencji dla zdrowia psychicznego.

- **Przetwarzanie danych wrażliwych, własności intelektualnej i danych osobowych,** w tej kategorii mieszczą się ryzyka związane z nielegalnym lub niepożądanym przetwarzaniem danych przez model językowy, co może skutkować ujawnieniem ich osobom niepowołanym.

- **Niejasna sytuacja prawna** związana z rozwijającymi się jeszcze regulacjami oraz niepewną linią orzeczniczą może rodzić niepewność w stosunku do niektórych zastosowań modeli językowych.

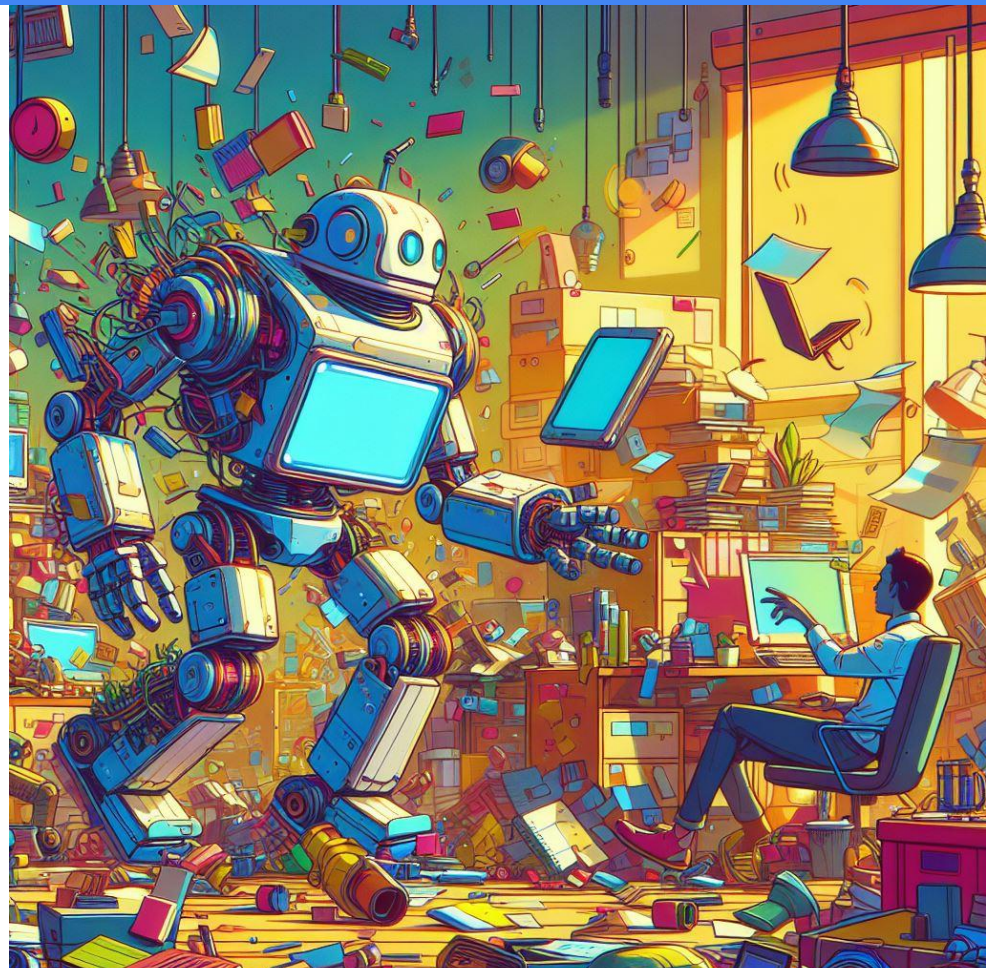
- **Brak ogólnoorganizacyjnych polityk dotyczących stosowania modeli językowych,** może spowodować problemy komunikacyjne oraz trudności w szkoleniu pracowników, a brak wiedzy o modelach językowych może prowadzić do błędnego korzystania z nich.

RYZYKA ZWIĄZANE ZE STOSOWANIEM MODELI JĘZYKOWYCH W PRACY

- **Różna jakość odpowiedzi w różnych językach,** to efekt trenowania modeli głównie na danych anglojęzycznych. Stąd wiedza na temat polskich danych jest mniejsza np. wiedza o polskim systemie prawnym. Dodatkowo prompty zadawane w języku angielskim dają z reguły wyższą jakość odpowiedzi i wiele osób korzysta z automatycznego tłumaczenia.

- **Trudność w ocenie nowych pracowników**, związana z możliwością wykorzystywania modeli językowych w procesie rekrutacji zarówno przez osobę rekrutowaną jak i osobę rekrutującą.

- **Nietransparentność** modeli językowych wynika z zastosowanej techniki uczenia, która tworzy sztuczną sieć neuronowa. Wyuczona na dużej ilości danych sieć nie dostarcza wyjaśnienia na temat sposobu swojego wnioskowania w języku zrozumiałym dla człowieka. Stąd, sieci neuronowe mogą posiadać nieznane dla nas właściwości.



RYZYKA ZWIĄZANE ZE STOSOWANIEM MODELI JĘZYKOWYCH W PRACY

Ryzyka techniczne (za OWASP)

- **Wstrzyknięcie komendy** polega na wprowadzeniu do systemu szkodliwych danych, które są interpretowane jako komendy. Może to prowadzić do nieautoryzowanego dostępu do systemu, manipulacji danymi lub wykonywania niepożądanych działań.

- **Niezabezpieczona obsługa danych wyjściowych**, występuje, gdy system nie zabezpiecza prawidłowo danych wyjściowych, co może prowadzić do ich nieautoryzowanego ujawnienia, manipulacji lub utraty.

- **Zatruwanie danych treningowych** polega na manipulacji danymi używanymi do trenowania modelu AI, co może prowadzić do nieprawidłowego działania systemu, błędów w prognozach lub decyzjach.

RYZYKA ZWIĄZANE ZE STOSOWANIEM MODELI JĘZYKOWYCH W PRACY

- **Odmowa usługi modelu** polega na przeciążeniu systemu AI zbyt dużą ilością żądań lub złożonymi danymi, co prowadzi do spowolnienia lub zatrzymania działania systemu.

- **Luki w zabezpieczeniach łańcucha dostaw** odnoszą się do słabości w procesach, narzędziach i technologiach używanych do tworzenia, dostarczania i utrzymania systemu AI, które mogą być wykorzystane do ataków.

- **Ujawnianie wrażliwych informacji** polega na nieautoryzowanym dostępie do poufnych danych przechowywanych, przetwarzanych lub transmitowanych przez system AI.

- **Kradzież modelu** polega na nieautoryzowanym dostępie i kopiowaniu modelu AI, co może prowadzić do utraty własności intelektualnej, niewłaściwego użycia modelu lub utraty przewagi konkurencyjnej.

- **Niebezpieczny projekt wtyczki** odnosi się do błędów projektowych lub implementacyjnych w wtyczkach używanych przez system AI, które mogą prowadzić do nieprawidłowego działania systemu, luk w zabezpieczeniach lub innych problemów.

- **Nadmierna agencja** polega na przyznaniu systemowi AI nadmiernych zdolności do podejmowania decyzji lub wykonywania działań, co może prowadzić do błędów, nieprawidłowości lub niepożądanych konsekwencji.

- **Nadmierna zależność** odnosi się do sytuacji, gdy organizacja lub system staje się zbyt zależny od konkretnego systemu AI, lub przestarzałych technik promptowania, co może prowadzić do problemów, jeśli system przestaje działać prawidłowo, jest niedostępny lub zostaje skompromitowany.

ZAGADNIENIA PRAWNE

Z perspektywy prawnej sytuacja wciąż daleko nam od dedykowanej regulacji. Ponad dwa lata temu Komisja Europejska (21 kwietnia 2021 r. z późniejszymi poprawkami) przedstawiła propozycję rozporządzenia ws. sztucznej inteligencji AI Act (AIA). Jej celem jest zapewnienie bezpiecznego rozwoju systemów AI, które nie naruszają praw podstawowych. Kolejny ważny akt pojawił się 28 września 2022 r. w którym to KE przedstawiła dyrektywę w sprawie pozaumownej odpowiedzialności cywilnej do sztucznej inteligencji (AILD).

Na dzień dzisiejszy nie mamy pewności:

- jakie zostaną przyjęte kluczowe definicje: sztucznej inteligencji, dostawców i operatorów oraz rozróżnienie rodzajów systemów
- jak zostanie ukształtowana odpowiedzialność za ew. szkodę jaką mogą ponieść uczestnicy rynku wynikającą z używania systemów AI.
- jak przepisy te zostaną wdrożone do polskiego porządku prawnego.

Co wynika z propozycji AIA?

Analizując kierunek w jakim podążają prace możemy jednak przypuszczać, że zakres regulacji będzie dość szeroki (tj. szeroka będzie definicja tego co ostatecznie uznane za sztuczną inteligencję). Niezbędne będzie dość świadome korzystanie z tych systemów i upewnianie się, że dostawcy (producenci) zapewniają zgodność z AIA. Planowane jest wprowadzenie specjalnego poziomu ochrony dla ewentualnych poszkodowanych decyzjami firm, które korzystać będą z AI. Poszkodowani będą na uprzywilejowanej pozycji w zakresie przedstawiania dowodów czy domniemania winy.

Rozporządzenie dzieli systemy AI na trzy kategorie ze względu na poziom ryzyka:

- systemy stwarzające niedozwolone ryzyko,
- systemy wysokiego ryzyka
- systemy ograniczonego ryzyka.

Systemy stwarzające niedozwolone ryzyko, takie jak te wykorzystujące techniki podprogowe czy identyfikację biometryczną w przestrzeni publicznej, mają zostać zakazane.

Systemy wysokiego ryzyka, takie jak te związane z identyfikacją biometryczną czy zarządzaniem infrastrukturą krytyczną, muszą spełniać szereg wymogów, a odpowiedzialność za ich funkcjonowanie spoczywa głównie na dostawcach systemów.

Systemy ograniczonego ryzyka to te, które nie należą do żadnej innej kategorii.

Wymogi Rozporządzenia AI Act dotyczą głównie systemów wysokiego ryzyka i obejmują sporządzenie odpowiedniej dokumentacji oraz zarządzanie danymi w sposób odpowiedni i bezpieczny. Wprowadzono także obowiązek rejestracji tego typu systemów w unijnej bazie danych."

ZAGADNIENIA PRAWNE

Zasady ogólne niezależne od klasyfikacji systemu:

- nadzór człowieka – rozwój i wykorzystywanie systemów AI powinno podlegać kontroli człowieka, w szczególności, kiedy decyzja algorytmiczna wpływa na kształtowanie praw;
- transparentność – możliwość śledzenia sposobu funkcjonowania systemu oraz odpowiedniego informowania o tym użytkowników;
- zarządzanie prywatnością – zgodność z zasadami ochrony prywatności i danych;
- niedyskryminacja, różnorodność i sprawiedliwość – systemy powinny powstawać z wykorzystaniem reprezentatywnych danych oraz nie mogą powielać uprzedzeń i prowadzić do dyskryminacji;
- bezpieczeństwo i niezawodność – minimalizowanie ryzyka wystąpienia nieoczekiwanych problemów;
- dobrostan społeczny i środowiskowy – zgodność z zasadami zrównoważonego i przyjaznego dla środowiska rozwoju.

Jak przygotować się nadchodzące regulacje?

- zacząć wprowadzać standardy korzystania z narzędzi AI, które będą mocno osadzone w ramach biznesowych, operacyjnych, biznesowych i prawnych aspektów działania firmy,
- wprowadzać szkolenia dedykowane rozumieniu działania AI i systemów AI przez pracowników, w szczególności zagrożeń i ryzyk jakie niesie ta technologia, aby byli bardziej świadomi podczas używania/testowania dostępnych na rynku narzędzi.
- zadbać, aby w firmie obecne były osoby z kompetencjami pozwalającymi stwierdzić jakie dane będą używane przez AI i na jakich ew. będą dotrenowywane systemu na potrzeby firmy, w szczególności zapewnienie zgodności z przepisami RODO oraz dbania o przestrzeganie wymogów stawianych przez AIA

Jeśli zainteresowały Cię tematy poruszane w tej broszurze, pozostań w kontakcie z Fundacją Optimum Pareto

- strona: optimumpareto.com
- email: fundacja@optimumpareto.com
- facebook: optimumparetofundacja
- autor broszury: Marcin Woźniak
  m.wozniak@optimumpareto.com

FUNDACJA OPTIMUM PARETO

Kraków

Zadanie publiczne dofinansowane ze środków Miasta Krakowa