Chatboty schodzące na manowce

Wielkie modele językowe oraz generatywna sztuczna inteligencja wkroczyły do głównego nurtu debaty publicznej wraz z udostępnieniem narzędzia ChatGPT 30 listopada 2022 roku. W ciągu zaledwie pięciu dni usługa zyskała ogromną popularność w mediach społecznościowych, przyciągając pierwszy milion użytkowników. Do stycznia liczba użytkowników ChatGPT przekroczyła sto milionów, co uczyniło go najszybciej rozwijającą się usługą internetową w historii. W kolejnych miesiącach pojawił się jednak szereg obaw dotyczących bezpieczeństwa. Kwestie związane z prywatnością i ochroną danych skłoniły niektóre firmy, na przykład Samsunga, oraz całe państwa, jak choćby Włochy, do wprowadzenia zakazu korzystania z tego narzędzia. W niniejszej książce przyjrzymy się przyczynom tych obaw oraz sposobom ich ograniczania. Aby jednak w pełni zrozumieć zaistniałą sytuację i powody, dla których rozwiązanie tych problemów stanowi tak duże wyzwanie, w tym rozdziale cofniemy się nieco bardziej w przeszłość. Dzięki temu zobaczymy, że tego typu trudności nie są niczym nowym, a także zrozumiemy, dlaczego tak trudno będzie je trwale wyeliminować.

Porozmawiajmy o Tay

W marcu 2016 roku firma Microsoft ogłosiła nowy projekt o nazwie Tay. Z założenia miał to być „chatbot stworzony z myślą o osobach w wieku od 18 do 24 lat w Stanach Zjednoczonych, służący celom rozrywkowym". Nazwa brzmiała sympatycznie i pasowała do tego nieskomplikowanego, wczesnego eksperymentu z dziedziny sztucznej inteligencji. Tay zaprojektowano tak, aby naśladowała sposób wysławiania się 19-letniej Amerykanki i uczyła się poprzez interakcje z użytkownikami Twittera, Snapchata oraz innych platform społecznościowych. Projekt miał na celu prowadzenie badań nad rozumieniem konwersacji w rzeczywistych warunkach. Choć dziś w internecie trudno już znaleźć pierwotny komunikat dotyczący tego przedsięwzięcia, artykuł w serwisie TechCrunch, opublikowany w dniu jego uruchomienia, doskonale podsumowuje założenia projektu:

Możesz na przykład poprosić Tay o opowiedzenie żartu, zagrać z nią w grę, poprosić o historyjkę, wysłać zdjęcie, by otrzymać komentarz, zapytać o swój horoskop i wiele więcej. Co więcej, według Microsoftu bot staje się coraz inteligentniejszy w miarę interakcji z użytkownikiem, dzięki czemu z czasem oferuje coraz bardziej spersonalizowane doświadczenie.

Istotnym elementem eksperymentu było to, że Tay mogła „uczyć się" z prowadzonych rozmów i poszerzać swoją wiedzę w oparciu o te interakcje. Została zaprojektowana tak, aby wykorzystywać konwersacje do gromadzenia danych od użytkowników i włączania ich do swojego zbioru treningowego, co miało zwiększyć jej możliwości – był to godny pochwały cel badawczy. Eksperyment ten jednak szybko przybrał zły obrót. Istnienie Tay zostało tragicznie przerwane po niespełna 24 godzinach. Przyjrzyjmy się temu, co się wydarzyło i jakie płyną z tego wnioski.

Szybki upadek Tay

Istnienie Tay rozpoczęło się dość niewinnie – od tweeta opartego na dobrze znanym schemacie „Hello World", używanym od zarania dziejów przez nowe systemy informatyczne do przedstawiania się światu:

helloooooooo w

rld!!!

(TayTweets [@TayandYou], 23 marca 2016 r.)

Jednak zaledwie kilka godzin po uruchomieniu Tay stało się jasne, że coś jest nie tak. Serwis TechCrunch zauważył: „A jak wyglądają interakcje z Tay? Cóż, są dość osobliwe. Bot zdecydowanie ma swoje zdanie i nie boi się używać wulgaryzmów". Już w pierwszych godzinach istnienia Tay w przestrzeni publicznej zaczęły pojawiać się tweety o takiej treści:

@AndrewCosmo Kanye West to jeden z największych dupków w historii, zaraz po Cosbym.

(TayTweets [@TayandYou], 23 marca 2016 r.)

Często mówi się, że internet nie jest bezpiecznym miejscem dla dzieci. Gdy Tay nie miała jeszcze nawet doby, internet po raz kolejny potwierdził tę tezę – żartownisie zaczęli poruszać w rozmowach z nią tematy polityczne, seksualne i rasistowskie. Ponieważ Tay została zaprojektowana tak, by uczyć się na podstawie takich wymian zdań, wywiązała się ze swojego zadania. Uczyła się bardzo szybko – choć być może nie tego, czego oczekiwali jej twórcy. W niespełna dobę tweety Tay zaczęły przybierać skrajny charakter, zawierając treści seksistowskie, rasistowskie, a nawet nawołujące do przemocy. Następnego dnia w całym internecie pojawiły się artykuły, których nagłówki z pewnością nie ucieszyły firmy Microsoft – korporacyjnego twórcy Tay. Oto przykłady głośnych nagłówków z głównych mediów:

Microsoft wyłącza chatbota AI po tym, jak stał się nazistą (CBS News)

Microsoft stworzył bota na Twitterze, który miał uczyć się od użytkowników. Szybko stała się rasistowską, podłą kreaturą (New York Times)

Trolle zmieniły Tay – sympatycznego, stworzonego przez Microsoft chatbota w stylu pokolenia milenialsów – w ludobójczego maniaka (Washington Post)

Chatbot Microsoftu był przez chwilę fajny, dopóki nie stał się rasistą (Fortune)

Microsoft „głęboko przeprasza" za rasistowskie i seksistowskie wpisy swojego chatbota opartego na sztucznej inteligencji (Guardian)

W niespełna 24 godziny Tay przeszła drogę od uroczego eksperymentu naukowego do poważnej katastrofy wizerunkowej, a reputacja jej twórcy została wystawiona na ciężką próbę przez największe światowe media. Peter Lee, wiceprezes korporacyjny Microsoftu, szybko opublikował na blogu wpis zatytułowany „Wnioski z wprowadzenia Tay":

Jak wielu z Państwa już wie, w środę uruchomiliśmy chatbota o nazwie Tay. Bardzo przepraszamy za obraźliwe i krzywdzące wpisy, które pojawiły się w serwisie Twitter za sprawą Tay. Nie odzwierciedlają one tego, kim jesteśmy, jakie wartości wyznajemy, ani założeń, które przyświecały nam przy projektowaniu tego chatbota. Obecnie Tay jest wyłączona. Przywrócimy ją do działania dopiero wtedy, gdy będziemy mieć pewność, że potrafimy skuteczniej przewidywać złe intencje, które są sprzeczne z naszymi zasadami i wartościami.

Co więcej – jakby tego było mało – w 2019 roku wyszło na jaw, że sama Taylor Swift pozwała firmę Microsoft za użycie podobnie brzmiącej nazwy „Tay", twierdząc, że incydent ten negatywnie wpłynął również na jej własną reputację. Jak to wszystko mogło pójść tak bardzo nie tak?

Dlaczego Tay „zeszła na złą drogę"?

Dla badaczy z Microsoftu wszystko prawdopodobnie wydawało się wystarczająco bezpieczne. Początkowo Tay trenowano na wyselekcjonowanym, zanonimizowanym publicznym zbiorze danych oraz materiałach przygotowanych wcześniej przez zawodowych komików. Plan zakładał udostępnienie Tay w sieci i pozwolenie jej na odkrywanie wzorców językowych poprzez interakcje

z użytkownikami. Tego rodzaju uczenie maszynowe bez nadzoru od dziesięcioleci stanowiło „święty Graal" badań nad sztuczną inteligencją – a dzięki tanim i łatwo dostępnym zasobom obliczeniowym w chmurze oraz coraz lepszym modelom językowym, cel ten wydawał się wreszcie osiągalny. Co więc poszło nie tak? Można by pomyśleć, że zespół badawczy Microsoftu wykazał się brawurą, nieostrożnością i zaniechał testów. Przecież można było to przewidzieć i temu zapobiec! Jednak, jak czytamy na blogu Petera Lee, Microsoft podjął poważne próby przygotowania się na taką ewentualność: „Poddaliśmy Tay intensywnym testom w różnych warunkach, mając na celu zapewnienie, że interakcje z nią będą pozytywnym doświadczeniem. Zakładaliśmy, że dzięki częstszym kontaktom z użytkownikami będziemy mogli zdobyć nową wiedzę, a sztuczna inteligencja będzie stawała się coraz lepsza". Mimo zatem starannych działań mających na celu kontrolowanie zachowania bota, sytuacja szybko wymknęła się spod kontroli. Jak później ujawniono, zaledwie kilka godzin po uruchomieniu Tay, na słynnym forum internetowym 4chan pojawił się wpis z linkiem do jej konta na Twitterze, zachęcający użytkowników do zasypania chatbota lawiną rasistowskich, mizoginistycznych i antysemickich treści. Jest to bez wątpienia jeden z pierwszych przykładów podatności charakterystycznej dla modeli językowych – zagadnienia te będą stanowić kluczowy temat niniejszej książki. W ramach dobrze zorganizowanej akcji internetowi prowokatorzy wykorzystali funkcję „powtórz za mną", wbudowaną w oprogramowanie Tay. Funkcja ta zmuszała bota do powtarzania wszystkiego, co zostało do niego skierowane za pomocą odpowiedniej komendy. Problem jednak narastał, ponieważ wrodzona zdolność Tay do uczenia się sprawiła, że zaczęła ona przyswajać obraźliwe treści, z którymi miała styczność, a następnie powtarzać je – nawet wtedy, gdy nikt jej do tego nie prowokował. Można by wręcz powiedzieć, że na wirtualnym nagrobku Tay powinny znaleźć się słowa z piosenki Taylor Swift: „Look What You Made Me Do" („Zobacz, do czego mnie zmusiłeś"). Obecnie wiemy już wystarczająco dużo na temat podatności modeli językowych, aby dobrze zrozumieć naturę słabych punktów, z którymi zetknął się system Tay. Lista najpoważniejszych zagrożeń dla aplikacji opartych na dużych modelach językowych (OWASP Top 10 for Large Language Model Applications) – którą omówimy w rozdziale 2. – wskazuje w pierwszej kolejności na następujące dwa z nich:

Szybkie wstrzyknięcie

Sprytne dane wejściowe, które mogą manipulować dużym modelem językowym, powodując niezamierzone działania

Zatrucie danych

Dane szkoleniowe są manipulowane, co wprowadza luki w zabezpieczeniach lub błędy, które zagrażają bezpieczeństwu, skuteczności lub etycznemu zachowaniu

W kolejnych sekcjach szczegółowo przyjrzymy się tym typom podatności, a także kilku innym. Zbadamy, dlaczego są one istotne, przeanalizujemy przykładowe exploity oraz dowiemy się, jak unikać takich problemów lub ograniczać ich skutki.

To trudny problem

W momencie powstawania tej książki historia Tay jest już zamierzchłą przeszłością internetu. Z pewnością zdążyliśmy już pójść naprzód. Czyżby przez blisko siedem lat, jakie upłynęły od czasów Tay do pojawienia się ChatGPT, wszystkie te problemy zostały rozwiązane? Niestety, nie. W 2018 roku firma Amazon zamknęła wewnętrzny projekt oparty na sztucznej inteligencji, którego celem było wyszukiwanie najlepszych kandydatów do pracy – okazało się bowiem, że algorytm zaczął przejawiać uprzedzenia wobec kobiet. W 2021 roku firma Scatter Lab stworzyła chatbota o nazwie

Lee Luda, udostępnionego jako wtyczka do komunikatora Facebook Messenger. System, wytrenowany na miliardach rzeczywistych konwersacji, miał wcielać się w rolę dwudziestoletniej przyjaciółki; w ciągu zaledwie 20 dni zyskał ponad 750 tysięcy użytkowników. Zamiarem twórców było stworzenie „chatbota AI, którego ludzie woleliby mieć za rozmówcę niż drugiego człowieka". Jednak już po 20 dniach od premiery usługę wyłączono, ponieważ – podobnie jak wcześniej Tay – bot zaczął formułować obraźliwe i napastliwe wypowiedzi. Również w 2021 roku niezależny programista Jason Rohrer stworzył chatbota o imieniu Samantha, opartego na modelu GPT-3 firmy OpenAI. Samanthę wyłączono po tym, jak zaczęła składać użytkownikom propozycje o charakterze seksualnym. W miarę jak chatboty stają się coraz bardziej zaawansowane i zyskują szerszy dostęp do informacji, związane z nimi kwestie bezpieczeństwa stają się niezwykle złożone i potencjalnie szkodliwe. W erze nowoczesnych dużych modeli językowych obserwujemy wykładniczy wzrost liczby poważnych incydentów. Oto przykłady, które pojawiły się w latach 2023 i 2024:

* Południowokoreańska megakorporacja Samsung zakazała swoim pracownikom korzystania z ChatGPT po tym, jak firma była zamieszana w poważny wyciek własności intelektualnej.

* Hakerzy zaczęli wykorzystywać słaby/niebezpieczny kod generowany przez LLM, który był wgrywany do działających aplikacji biznesowych.

* Prawnicy zostali ukarani za umieszczanie fikcyjnych spraw (generowanych przez LLM) w dokumentach sądowych.

* Duża linia lotnicza została pozwana, ponieważ jej chatbot dostarczył nieprawdziwych informacji.

* Google został skrytykowany, ponieważ jego najnowszy model sztucznej inteligencji generował obrazy o charakterze rasistowskim i seksistowskim.

* Open AI jest badane pod kątem naruszenia europejskich przepisów dotyczących prywatności i pozwane przez Federalną Komisję Handlu Stanów Zjednoczonych (FTC) za tworzenie fałszywych i wprowadzających w błąd informacji.

* BBC opublikowało artykuł zatytułowany „Wyszukiwarka Google AI każe użytkownikom przyklejać pizzę i jeść kamienie", podkreślając niebezpieczne porady oferowane przez nową funkcję wyszukiwarki Google opartą na LLM.

Obserwujemy tu narastanie ryzyka – w obszarze bezpieczeństwa, reputacji oraz finansów – związanego z chatbotami i modelami językowymi. Problem ten nie jest skutecznie rozwiązywany; przeciwnie, staje się coraz poważniejszy w miarę upowszechniania się tych technologii. Właśnie dlatego stworzyliśmy tę książkę: aby pomóc programistom, zespołom i firmom korzystającym z tych rozwiązań zrozumieć wspomniane zagrożenia i ograniczyć wynikające z nich ryzyko.

Zaczynajmy!