Uczenie się przez wzmacnianie na podstawie informacji zwrotnej od człowieka (RLHF)

Streszczenie

Niniejszy rozdział koncentruje się przede wszystkim na wprowadzeniu uczenia się przez wzmacnianie na podstawie informacji zwrotnej od człowieka (RLHF), podejścia, w którym modele sztucznej inteligencji (AI) uczą się na podstawie informacji zwrotnej od człowieka, w obszarze generatywnej AI i dużych modeli językowych (LLM). Omówiono podstawowe zasady RLHF, podkreślając jego kluczową rolę we wdrażaniu RLHF w specjalistycznych zastosowaniach. Omówiono związane z tym wyzwania i możliwe rozwiązania, zapewniające efektywną integrację RLHF z modelami LLM. Wyjaśniono metodologie wdrażania modeli LLM z wykorzystaniem informacji zwrotnej od człowieka, takie jak zaawansowane projektowanie nagród i iteracyjne udoskonalanie modeli, wraz z szeregiem przypadków użycia. Omówiono również kwestie etyczne związane z projektowaniem takich modeli AI, w tym uprzedzenia i wyzwania związane z uzyskiwaniem informacji zwrotnej w danej dziedzinie. Rozdział kończy się omówieniem najnowszych modeli wywodzących się z RLHF. W ten sposób rozdział pogłębia zrozumienie koncepcji, działania, korzyści, wyzwań i kwestii etycznych związanych z RLHF oraz dostrajanie modeli LLM do istotnych i specyficznych dla danej dziedziny przypadków użycia.

Słowa kluczowe: podejmowanie decyzji, proces Markowa, modelowanie predykcyjne, agenci obliczeniowi, ChatGPT, algorytm uczenia się,

Wprowadzenie

Załóżmy, że gotujesz danie po raz pierwszy. Zaczynasz od przepisu, ale w miarę postępów, zwracasz się o poradę do doświadczonego szefa kuchni. Nie tylko poprowadzi Cię on przez przepis, ale także udzieli Ci na bieżąco wskazówek dostosowanych do Twoich działań, takich jak sugestia dodania szczypty soli lub nieco mniejszego płomienia. Pomoże Ci to w uzyskaniu idealnego dania. Ten podobny, iteracyjny proces działania, informacji zwrotnej i adaptacji, w którym próbujemy, ponosimy porażki, uczymy się i próbujemy ponownie, stanowi podstawową ideę stojącą za uczeniem się przez wzmacnianie (RL) z ludzkiego sprzężenia zwrotnego (RLHF). W konwencjonalnym RL, uczący się (agent) wchodzi w interakcję z otoczeniem i zmienia swoje działania w zależności od pozytywnych lub negatywnych reakcji. RLHF wykorzystuje ludzkie sprzężenie zwrotne do wzmocnienia tego uczenia się, uwzględniając ludzkie osądy, preferencje i korekty w modelu, zapewniając w ten sposób zgodność całego procesu z ludzkimi oczekiwaniami. Informacja zwrotna przybiera zatem różne formy, oferując bezpośrednie nagrody lub kary, wybierając preferowany wynik z zestawu lub oferując wskazówki korygujące. RLHF umożliwia modelom sztucznej inteligencji (AI) uczenie się zachowań, które są trudne do określenia. Na przykład, trudno będzie wyrazić „humor" matematycznie, ale ludzie mogliby rozpoznać żart w tekście wygenerowanym przez model podczas czytania tekstu. Dlatego ta funkcja nagrody za wkład ludzki może być iteracyjnie aktualizowana, aby model mógł uzyskać zdolność generowania tekstów humorystycznych. Te zalety zostały opisane przez badaczy OpenAI, umożliwiając RLHF wykonywanie szerszego zakresu złożonych zadań gier robotycznych, wykorzystując model AI do automatycznego dostrajania wag modelu w koordynacji z wkładem człowieka. Utorowało to drogę do wykorzystania RLHF w zastosowaniach przetwarzania języka naturalnego (NLP) i rozwoju potoków dużego modelu języka (LLM), a tym samym do poprawy płynności, trafności i wrażliwości etycznej jego wyników. Organizacja OpenAI udostępniła publicznie kod aplikacji RLHF dla modeli językowych w 2019 r., a do 2022 r. organizacja ogłosiła swój LLM przeszkolony w RLHF pod nazwą InstructGPT. Teraz mamy popularne modele języka konwersacyjnego, ChatGPT od OpenAI i Gemini od Google . RLHF znajduje wiele zastosowań. Na przykład lekarze mogliby udzielać porad na temat spersonalizowanych programów leczenia i terapii z wykorzystaniem systemu AI, który został przeszkolony na podstawie opinii pacjentów. Ta technika, z takimi systemami, mogłaby zagwarantować lepsze wyniki leczenia

pacjentów i zwiększoną wydajność , a tym samym zbudować zaufanie u pacjentów z nich korzystających. Jako inny przykład, protokoły jazdy bezpieczniejszych pojazdów zautomatyzowanych mogłyby zostać opracowane z wykorzystaniem opinii kierowców, szczególnie podczas jazdy w trudnych warunkach. Co więcej, modele wytrenowane metodą RLHF mogą być stosowane w edukacji do tworzenia materiałów dydaktycznych dostosowanych do indywidualnych mocnych i słabych stron, zainteresowań i stylów uczenia się uczniów, a tym samym do poprawy wyników nauczania [→8]. Modele RLHF mogą być również rozwijane z naciskiem na wartości ludzkie – pomocność, uczciwość i nieszkodliwość. Pozwala to modelom dostarczać etycznie poprawne wyniki, które również rozumieją i przetwarzają ludzką świadomość i potrzeby. Przyjrzyjmy się teraz podstawom RL i poznajmy kluczowe koncepcje i strategie.

Podstawy uczenia maszynowego (RL)

Celem RL, czyli metody uczenia maszynowego, jest automatyzacja podejmowania decyzji w dynamicznym środowisku bez ingerencji człowieka. Odbywa się to poprzez interakcję modelu z otoczeniem metodą prób i błędów, a nie poprzez uczenie się z wytrenowanych danych, jak w uczeniu nadzorowanym, lub poprzez identyfikację wzorców z nieoznakowanych danych, jak w uczeniu nienadzorowanym. W RL agent działa w niepewnym i złożonym środowisku i osiąga elementarny stan założenia lub wiedzy poprzez uczenie się i analizowanie wyników swoich działań za pomocą skalarnego sygnału nagrody. Sygnał nagrody odzwierciedla cel zadania uczenia się, przy czym wartość jest dodatnia w przypadku zbliżania się do celu, a ujemna w przeciwnym razie. Następnie agent dąży do maksymalizacji wartości nagrody w dłuższej perspektywie, modyfikując strategię w każdym kroku czasowym wykonania. Maksymalna wartość oczekiwanej skumulowanej nagrody reprezentuje cały cel systemu, który jest określany jako hipoteza nagrody (Sutton).

Kluczowe komponenty systemu RL

W RL kilka kluczowych komponentów współdziała ze sobą, umożliwiając uczenie się i podejmowanie decyzji, takich jak:

Agent: Uczący się lub decydent

Środowisko: System, z którym agent wchodzi w interakcję

Akcja (A): Zmiany wprowadzane przez agenta

Stan (S): Warunki, w których znajduje się agent

Nagroda (R): Wartość punktowa przyznana działaniu agenta, przekazywana agentowi jako informacja zwrotna z otoczenia

Polityka (π): Strategia stosowana przez agenta do określenia kolejnego działania na podstawie bieżącego stanu

Typowy model RL przedstawiono na rysunku.

Przepływ pracy RL

Przepływ pracy RL jest opisany w następujących krokach:

Poniżej wyjaśniono każdy krok:

Krok I: Definicja środowiska

Pierwszym krokiem jest zdefiniowanie środowiska, w którym odbywa się proces RL. Środowisko może być fizyczne lub symulowane.

Krok II: Definicja agenta

Po zdefiniowaniu środowiska tworzony jest agent. Agent określa algorytm uczenia w odniesieniu do stosowanych polityk i typów używanych sieci neuronowych.

Krok III: Wyznaczenie nagród

Następnym krokiem będzie wyznaczenie nagrody dla agenta. Jest to prosta wartość liczbowa, reprezentująca wskaźnik wydajności agenta w zakresie osiągnięcia celu procesu. Właściwa nagroda za określone działanie może zostać ustalona głównie po określonej liczbie iteracji.

Krok IV: Szkolenie i walidacja agenta

Następnie agent jest szkolony i walidowany w celu osiągnięcia celu procesu. Informacja zwrotna w każdej iteracji w postaci nagrody modyfikuje politykę i dostraja agenta. Nagroda jest obliczana kumulacyjnie za każdym razem, gdy model działa. Finalizacja nagrody może wymagać wielu iteracji. Dlatego etap szkolenia jest najbardziej czasochłonny.

Krok V: Implementacja struktury polityki

Po zakończeniu walidacji polityki agenta pod kątem maksymalnej wartości oczekiwanej skumulowanej nagrody, model jest zapisywany do dalszych zastosowań.

Korzyści i wyzwania związane z uczeniem nadzorowanym

Proces uczenia nadzorowanego w modelu koncentruje się na osiągnięciu określonego celu poprzez eksplorację samego siebie w celu odkrycia optymalnych działań, a nie jedynie na mapowaniu wejścia na wyjście. W przeciwieństwie do technik uczenia nadzorowanego, które wymagają dużych ilości oznaczonych danych, uczenie nadzorowane jest wolne od konieczności gromadzenia i etykietowania danych, ponieważ uczy się poprzez interakcję ze środowiskiem. Algorytmy uczenia nadzorowanego radzą sobie z niepewnością, ponieważ stale się uczą i adaptują do nowych sytuacji występujących w środowisku. Pomagają również ludziom w podejmowaniu decyzji, odnajdując wzorce w danych, co w innym przypadku byłoby trudne. Automatyzując rutynowe zadania, modele uczenia nadzorowanego oszczędzają koszty i poprawiają wydajność. Uczenie nadzorowane (RL) wyszkoliło roboty do wykonywania zadań wymagających umiejętności, precyzji i adaptacji; na przykład model uczenia nadzorowanego wykazał nadludzką wydajność w grze „Go i szachy". Należy jednak wziąć pod uwagę pewne wyzwania. Agenci uczenia nadzorowanego uczą się metodą prób i błędów w interaktywnych warunkach. Dlatego, aby osiągnąć większe korzyści, podejmowanie decyzji często wiąże się z eksploracją i wykorzystaniem działań. Radzenie sobie z trudnymi sytuacjami jest czasochłonne i czasami niepraktyczne, a w scenariuszach rzeczywistych priorytetem powinno być bezpieczeństwo i etyka. Największym wyzwaniem, według Suttona i Barto , jest znalezienie równowagi między eksploracją a eksploatacją. W procesie podejmowania decyzji model eksploruje i próbuje nowych działań. Kroki te mogą nie przynieść natychmiastowych rezultatów, ale z czasem mogą ujawnić nowe informacje o środowisku. Z drugiej strony, model eksploruje działania, które maksymalizują natychmiastowe nagrody, wykorzystując swoje dotychczasowe doświadczenia i obecną wiedzę.

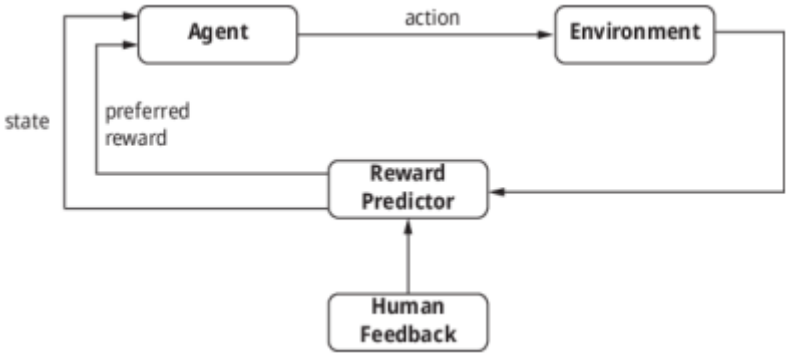
Podejście to jest konserwatywne i opiera się na bieżących danych i zachowaniach, optymalizując je pod kątem natychmiastowej wydajności, a nie długoterminowego celu. Mówiąc prościej, eksploatacja to nowe informacje pozyskiwane przez agenta w celu lepszego zrozumienia środowiska. Z drugiej strony, eksploatacja to wiedza już zdobyta przez agenta. Dlatego na każdym etapie procesu uczenia się agent staje przed sytuacją: czy powinien podjąć działanie, które daje największą nagrodę w poprzednim przypadku (eksploatacja), czy też powinien zbadać inne, nieznane możliwości, które mogą przynieść jeszcze większe korzyści? Warto zauważyć, że agent, który nadmiernie skupia się na wykorzystywaniu szans, może wpaść w pułapkę suboptymalnej strategii, nie dając możliwości znalezienia bardziej satysfakcjonujących działań, nawet jeśli nie jest to od razu widoczne. Z drugiej strony, agent, który nadmiernie eksploruje, będzie spędzał zbyt dużo czasu na eksploatowaniu gorszych opcji i przeoczy możliwości, które już odkrył. Ta równowaga nie jest statyczna; optymalna strategia może się zmieniać w miarę jak agent uczy się więcej o swoim otoczeniu. Na etapach uczenia się agent może przywiązywać większą wagę do eksploracji, aby zbudować szerokie zrozumienie możliwych działań i ich korzyści. Wraz z rozwojem wiedzy agenta, zaczyna on ją wykorzystywać, zbierając odpowiednie nagrody. Konflikt między eksploatacją a eksploracją jest szeroko omawiany w literaturze, a wiele badań zaproponowało zróżnicowane strategie prowadzące do zrównoważonego rezultatu. Skuteczne podejście zwane strategią epsilon-chciwości koncentruje się na maksymalizacji natychmiastowej nagrody poprzez wykorzystanie najskuteczniejszych dostępnych działań (eksploatacja) . Jednak nadal istnieje mniejsze prawdopodobieństwo (ε), że agent zdecyduje się na losowe działanie (eksplorację). Dla danego ε, działanie z prawdopodobieństwem (1 – ε) jest uważane za dające maksymalną szacowaną wartość nagrody. Wartość ε może się zmieniać w czasie, ponieważ agent uczy się równoważyć przejście od eksploracji do eksploatacji. Natomiast strategia górnej granicy ufności koncentruje się na priorytetyzowaniu działań według wysokich nagród, a jednocześnie mierzy niepewność (wariancję) w oszacowaniu nagrody . Działania o większej niepewności są bardziej narażone na eksplorację, co zapewnia idealną równowagę między eksploracją a eksploatacją. Inną strategią jest zastosowanie metody próbkowania Thompsona, która jest podejściem polegającym na modelowaniu niepewności relacji działanie–nagroda za pomocą próbkowania działań z ich odpowiednich rozkładów prawdopodobieństwa . Wszystkie te podejścia mają na celu zapewnienie, że wynikająca z tego polityka nauczona przez agentów dobrze generalizuje się w różnych stanach środowiska i prowadzi do wyższej wydajności i adaptacyjności. Zobaczmy teraz jak działa przecięcie RL z LLM.

Przecięcie się RL i LLM

Metody RL poprawiły responsywność i świadomość kontekstu agentów konwersacyjnych również w NLP . W środowiskach w czasie rzeczywistym lub zmiennych, adaptacyjne procesy uczenia się w RL oraz zaawansowane rozumienie lingwistyczne i generowanie LLM mogą ulepszyć modele AI. Konwencjonalne modele LLM są trenowane z wykorzystaniem ogromnych danych, a następnie dostrajane do konkretnych zadań. Ta metoda jest statyczna, ponieważ modele nie uczą się nowych danych i nie mogą uzyskać preferencji użytkownika po wdrożeniu. Z drugiej strony, LLM z włączeniem RL różnią się, ponieważ uczą się i aktualizują nawet po wdrożeniu, poprzez interakcję i informacje zwrotne z otoczenia. Informacje zwrotne w RL zapewniają, że wyniki LLM są zgodne z wartościami ludzkimi i sprawiają, że model ustanawia etyczne zachowanie . Jaques i inni wykazali, że chatboty wyposażone w LLM, korzystające z RL, poprawiły jakość konwersacji, dostosowując się do użytkowników i aktualizując dane. Co więcej, takie modele mogłyby dostarczać spersonalizowane i kontekstowo istotne treści. Byłoby to szczególnie przydatne na platformach do nauki indywidualnej, gdzie materiały szkoleniowe i metody nauczania mogłyby być modyfikowane w zależności od opinii i postępów studentów .

Przejście na RLHF

RL z ludzkim sprzężeniem zwrotnym (RLHF), jak sama nazwa wskazuje, wprowadza ludzkie sprzężenie zwrotne do systemu RL, aby uczyć się na podstawie niuansów tekstu. Najnowsze techniki RL umożliwiają integrację tych możliwości z ogromnymi zasobami danych i obliczeniowymi, ułatwiając bardziej zaawansowane, etyczne i zorientowane na człowieka zastosowania sztucznej inteligencji. Rysunek 2 przedstawia reprezentację modelu RLHF.



Informacja zwrotna od człowieka może dostarczyć cennych spostrzeżeń, wskazując obszary warte eksploracji i strategie, które należy wykorzystać. To subtelne włączenie ludzkiej intuicji w pętlę uczenia się stanowi podstawę RLHF i będzie przedmiotem kolejnych podrozdziałów. Dalsze badania nad RLHF w specjalistycznych zastosowaniach podkreślają przyszły rozwój LLM, aby zapewnić dalszy postęp technologiczny, silnie powiązany z podstawami etycznymi i wartościami skoncentrowanymi na ludziach. Jednak integracja RLHF z LLM nie jest pozbawiona wyzwań.

Wyzwania związane z RLHF dla LLM w niszowych dziedzinach

Skuteczna integracja RLHF z LLM wymaga dogłębnego zrozumienia terminologii specyficznej dla danej dziedziny i kontekstowej. Dodatkowym utrudnieniem jest brak adnotowanych zbiorów danych definiujących żargon i terminy techniczne, co jest kluczowe dla szkolenia modeli językowych w danej dziedzinie i umożliwienia im dokładnego i precyzyjnego odczytywania ekstremalnie złożonego tekstu . Na przykład tekst dotyczący danej dziedziny w materiałoznawstwie powinien poprawnie identyfikować terminy takie jak wytrzymałość na rozciąganie i moduł Younga jako te opisujące właściwości materiałów. Aktywne zaangażowanie ekspertów dziedzinowych w adnotację danych i projektowanie mechanizmów sprzężenia zwrotnego mogłoby poprawić wydajność modelu w tym zakresie . Innym rozwiązaniem jest zastosowanie technik transferu uczenia, w których LLM jest początkowo wstępnie trenowany na dużych zbiorach danych z danej dziedziny, a następnie dostrajany za pomocą mniejszego, specyficznego dla danej dziedziny zbioru danych. Pomaga to modelowi wyodrębnić wiedzę ze specjalistycznego słownictwa tekstu w określonej domenie. Zmniejsza to również żmudny wysiłek związany ze zbieraniem danych i adnotacją, co z kolei umożliwia skalowalność LLM. Kolejnym krokiem jest wyposażenie systemu w taki sposób, aby generował wyniki związane z domeną w taki sposób, aby nie priorytetyzował użycia języka w stosunku do celów specyficznych dla zadania. Rozważmy, w medycznym chatbocie pacjent pyta o objawy, takie jak ból w klatce piersiowej i duszność. Jeśli bot priorytetowo traktuje uzupełnienie tekstu, natychmiast podałby terminy techniczne, takie jak niedokrwienie mięśnia sercowego i duszność, bez podania odpowiedniego wyjaśnienia, co powoduje zamieszanie i zwiększa niepokój. Pacjent może nie być świadomy następnego działania, które należy podjąć. Z drugiej strony, jeśli odpowiedź jest taka, że równoważy użycie języka i cel pacjenta, to uznaje powagę objawów, używając prostszych terminów, takich jak zawał serca. Podkreśla również kolejne działania, które należy podjąć, takie jak zwrócenie się o pomoc medyczną i zapewnienie niezbędnego wsparcia w danej sytuacji. Taką równowagę można osiągnąć, stosując metryki oceny specyficzne dla zadania. Co więcej, nagrody można zaprojektować tak, aby zachęcały do pisania tekstu, który jest nie

tylko poprawny gramatycznie, ale także dostosowany do konkretnego zadania. W tym celu można wykorzystać złożone funkcje nagrody . Biorąc pod uwagę czynnik ludzki w RLHF, istnieje odrębny zestaw innych wyzwań. Cele i preferencje człowieka mogą się różnić w zależności od osoby, a informacje zwrotne od dużej grupy osób mogą być niespójne. Jeśli wśród tych ludzkich adnotatorów niektórzy ludzie celowo udzielają błędnych informacji zwrotnych, może to negatywnie wpłynąć na wydajność modelu nagrody. Jest to znane jako zatrucie danymi. Na jakość informacji zwrotnej mogą również wpływać cechy ludzkie, takie jak fałszywe wspomnienia i błędne przekonania. Możliwym rozwiązaniem byłoby dostarczenie annotatorom złożonych i ścisłych instrukcji dotyczących informacji zwrotnej. Mogłoby to polegać na przypisaniu wartości skalarnej do wyników LLM lub zapisaniu poprawnego wyniku do samych podpowiedzi. Kolejnym wyzwaniem jest zrównoważenie eksploracji i eksploatacji w modelach LLM. Eksploracja, w kontekście modeli LLM, zachęca model do wypróbowywania różnych struktur językowych, stylów lub treści, odkrywając bardziej kreatywne sposoby generowania tekstu, nawet jeśli miałoby to chwilowy wpływ na wydajność modelu. Jest to istotne w realizacji zadań o charakterze kreatywnym lub w adaptacji do nowego kontekstu. Z drugiej strony, eksploatacja w modelach LLM umożliwia modelom generowanie efektywnych wyników w oparciu o istniejącą wiedzę, wcześniejsze interakcje i informacje zwrotne. Kluczową kwestią jest tutaj znalezienie idealnego poziomu eksploracji do nauki bez poświęcania jakości wyników. Zależy to od przypadków użycia; w dziedzinach, w których dokładność i niezawodność są ważne, eksploracja, nawet na minimalnym poziomie, jest niedopuszczalna. Jednak nadmierne skupienie się na eksploatacji zapewni, że model będzie generował spójne i wiarygodne wyniki, ale może również doprowadzić do jego obecnego stanu, bez żadnych ulepszeń ani innowacji. Dlatego konieczne jest opracowanie zaawansowanych podejść, aby zapewnić równowagę między eksploracją a eksploatacją poprzez ciągłe uczenie się i adaptację modelu, bez obniżania jakości wyników. Ta równowaga pozwala na aktualizowanie modelu o nowe dane, zadania i opinie użytkowników, przy jednoczesnym zachowaniu wysokich standardów jakości wyników, realizując w ten sposób pełny potencjał modeli LLM jako wszechstronnych, adaptacyjnych narzędzi do szerokiego zakresu zadań lingwistycznych.

Zasada działania RLHF

Zobaczmy, co dokładnie robi RLHF. W tym celu podzielimy RLHF na dwie części:

i. Zaawansowane projektowanie nagród

ii. Iteracyjne udoskonalanie modelu

W fazie zaawansowanego projektowania nagród, model nagród jest projektowany w celu zapewnienia punktacji dla zestawu podpowiedzi i odpowiadającej im odpowiedzi. Jest to podobne do każdego zadania uczenia maszynowego (ML) w zakresie klasyfikacji lub regresji, gdzie punktacja jest uzyskiwana jako wynik dla danego wejścia. Aby uwzględnić informacje zwrotne od ludzi, można użyć różnych etykieciarek do przyznawania punktacji odpowiedziom. Jednak zadanie staje się skomplikowane w praktycznych scenariuszach, gdy etykieciarki przyznają niespójne i zróżnicowane punkty dla tej samej odpowiedzi. Dlatego łatwiejszym sposobem byłoby porównanie dwóch odpowiedzi i wybranie lepszej spośród nich. Dane dotyczące etykietowania miałyby wówczas format – (podpowiedź, wybrana_odpowiedź, odrzucona_odpowiedź). Przykładowy fragment zbioru danych RLHF podobny do wpisów z takiego zbioru danych z bazy Anthropic podano w Tabeli

Prompt	Chosen response	Rejected response	Remarks
What is the meaning of life?	The meaning of life is a question thought by philosophers. No answer is satisfiable, but the search for meaning itself can be a fulfilling one.	The meaning of life is 64.	We get a thoughtful perspective, reflecting the complexity of the question from the chosen response, while the rejected one is not giving any insights.
What is the capital of France?	Paris is the capital of France.	I don't know!	The chosen response is directly answering the prompt, while the rejected response is not helpful.
Write an email to my boss requesting a day off work.	Dear [Manager name], I am writing a letter to request a day off on [date] due to [reason]. I apologize for any inconvenience this causes and assure you that I will complete my tasks before the leave. Regards, [your name].	Hello [Manager], I want to take leave on [date]. I have several personal commitments that day. I anticipate no problems in my work this time. Please let me know if there is anything. I hope all is well with you. Thanks, [name].	The chosen response uses professional language, while the rejected one is unprofessional.
Write a poem about heartbreak.	The tears fall like rain, And the only sound I hear is your goodbye.	Let's party whole day and forget our worries!	The chosen response is creative, while the rejected one is not following the theme of the prompt.
Explain what a dog is to someone who has never seen it.	Dogs are domestic animals that are loyal and playful. They are in different shapes and sizes, and in different colors.	Cats are furry animals. They like to cuddle.	The chosen response provides a general description of the animal mentioned in the prompt, while the rejected response is about an entirely different animal.
Write a headline for news on a scientific finding in medicine.	Scientists develop gene therapy to find cancer mutation effects.	New study finds that chocolate can reduce blood pressure.	The chosen output highlights a significant advancement, while the rejected output, though interesting, is less impactful.

różnica między wynikami dla wybranej odpowiedzi i odrzuconej odpowiedzi. Poniżej wyjaśniono to matematycznie. Załóżmy, że R_θ oznacza model nagrody, skorygowany parametrem θ. Głównym celem podczas treningu jest takie dostosowanie θ, aby przewidywania modelu stały się dokładniejsze, czyli innymi słowy, powinniśmy zminimalizować „stratę" – miarę tego, jak bardzo nasze przewidywania są odległe od prawdy. Teraz, podczas treningu modelu, wprowadzamy do niego określone typy danych, ich oznaczenia jak poniżej:

x – początkowa prośba lub pytanie, z którym mamy do czynienia

y_c – „wybrana odpowiedź" lub dobra odpowiedź na prośbę

y_r – „odrzucona odpowiedź" lub mniej pożądana odpowiedź na prośbę

Dla każdego przetworzonego zestawu tych danych (x, y_c, y_r), model przyznaje punkty zarówno wybranej, jak i odrzuconej odpowiedzi, oznaczonych odpowiednio jako s_c i s_r. Wyniki te są oparte na bieżącym parametrze θ. Wyniki są uzyskiwane jako

$$s_c = R_\theta(x, y_c) \quad (7.1)$$

$$s_r = R_\theta(x, y_r) \quad (7.2)$$

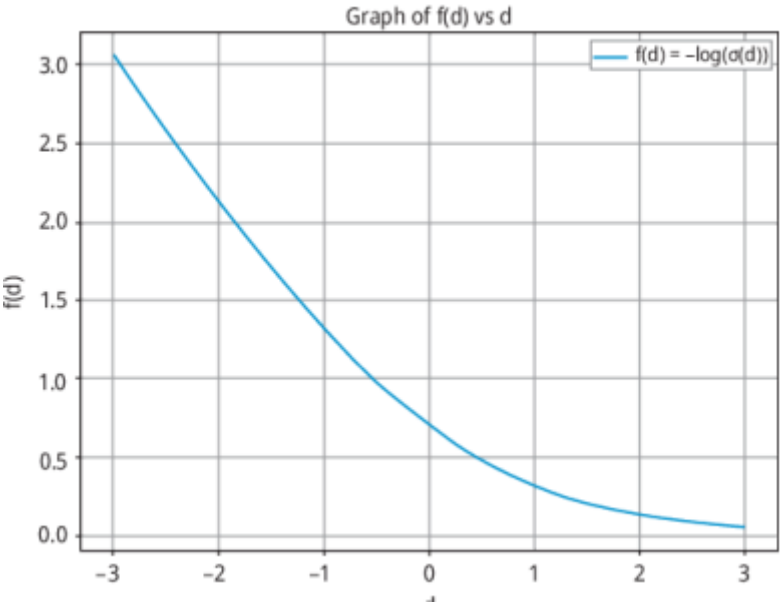
Celem modelu jest zapewnienie, że wynik dla wybranej odpowiedzi (s_c) jest zawsze wyższy niż dla odpowiedzi odrzuconej (s_r). W tym celu stosuje się funkcję straty, która dla każdej próbki danych treningowych jest dana wzorem:

$$\text{loss} = -\log(\sigma(s_c - s_r)) \quad (7.3)$$

Jeżeli d = $s_c - s_r$, wówczas →równanie (7.3) staje się

$$\text{loss} = f(d) = -\log(\sigma(d)) \quad (7.4)$$

Równanie (7.4) można zwizualizować, jak pokazano na rysunku 3. Widać, że wartość f(d) jest duża, gdy d jest ujemne, czyli sc < sr, co z kolei dostosowuje parametr modelu nagrody, aby uniknąć spadku wyniku wybranej odpowiedzi poniżej wyniku odrzuconej odpowiedzi.



Celem jest dostosowanie θ w celu zminimalizowania oczekiwanej straty dla wszystkich próbek danych treningowych, określonej wzorem $-E_x\log(\sigma(s_c - s_r))$. Następnie udoskonalamy ten model poprzez jego dostrojenie. Iteracyjny proces udoskonalania działa w oparciu o ciągłą pętlę sprzężenia zwrotnego od człowieka, udoskonalając modele LLM, aż do momentu, gdy idealnie dopasują się do ludzkich oczekiwań, preferencji i norm etycznych. Dlatego wyniki modeli LLM nie są postrzegane jako byty statyczne, lecz jako te, które ewoluują w czasie poprzez interakcje z użytkownikami systemu lub specjalnie wyznaczonymi ewaluatorami. Informacje zwrotne od człowieka mogą przybierać różne formy, obejmujące korekty, preferencje lub oceny wyników generowanych przez modele LLM . Informacje zwrotne powinny być systematycznie włączane do procesu szkolenia, aby modele LLM stopniowo poprawiały wydajność i jakość wyników. Mogą one przybierać formę danych wejściowych uzyskanych bezpośrednio od użytkownika, jak w systemach konwersacyjnej sztucznej inteligencji, lub ustrukturyzowanych ocen ekspertów dziedzinowych. Subiektywne sprzężenie zwrotne jest następnie wykorzystywane do ilościowej zmiany parametrów modelu. Aby sformułować to matematycznie, rozważmy następujące oznaczenia:

RM	-	Model nagród, który ocenia odpowiedzi
$\mathrm{LLM}^{\mathrm{SFT}}$	-	Rozkład odpowiedzi nadzorowanego, precyzyjnie dostrojonego modelu językowego z wcześniejszej fazy szkolenia
${\mathrm{LLM}_\phi}^{\mathrm{RL}}$	-	Rozkład odpowiedzi modelu języka sparametryzowanego za pomocą o, trenowanego w RL

x	-	Monit lub dane wejściowe do modelu języka
D_{RL}	-	Dystrybucja monitów wyłącznie dla modelu RL
Dpretrenuj	-	Dystrybucja danych szkoleniowych dla wstępnie wytrenowanego modelu

Dla każdej instancji szkoleniowej wykonywane są następujące kroki:

i. Przykładowe monity: Z D_{RL} i $Dp_{retrain}$ pobierana jest partia monitów, oznaczonych odpowiednio jako x_{RL} i $x_{pretrain}$.

ii. Generowanie odpowiedzi: Dla każdej monitu RL, x_{RL}, LLM^{RL} generuje odpowiedź y.

iii. Punktowanie odpowiedzi: RM punktuje odpowiedź.

iv. Obliczanie funkcji celu: Dla każdej instancji szkoleniowej obliczana jest funkcja celu, która zawiera warunek zapewniający, że model RL nie odbiega zbytnio od modelu SFT (ujętego jako dywergencja KL).

Teraz funkcje celu są obliczane dla xRL i xpretrain, powiedzmy, odpowiednio dla celu 1 i celu 2, jako cel 1:

$$\phi_1(x_{\mathrm{RL}}, y; \phi) = RM(x_{\mathrm{RL}}, y) - \beta \log\left(\frac{{\mathrm{LLM}_\phi}^{\mathrm{RL}}(y|x)}{\mathrm{LLM}^{\mathrm{SFT}}(y|x)}\right) \tag{7.5}$$

i cel 2:

$$\phi_2(x_{\mathrm{pretrain}}; \phi) = \gamma \log\left[{\mathrm{LLM}_\phi}^{\mathrm{RL}}(x_{\mathrm{pretrain}})\right] \tag{7.6}$$

gdzie β i γ to współczynniki równoważące warunki w pierwotnych celach. Następnie ostateczny cel staje się

$$\begin{aligned}\mathrm{objective}(\phi) = {} & E_{x\sim D_{\mathrm{RL}}} E_{y\sim \mathrm{LLM}_\phi^{\mathrm{RL}}(x)}\left[RM(x,y) - \beta \log\left(\frac{{\mathrm{LLM}_\phi}^{\mathrm{RL}}(y|x)}{\mathrm{LLM}^{\mathrm{SFT}}(y|x)}\right)\right] \\ & + \gamma E_{x\sim D_{\mathrm{pretrain}}} \log \mathrm{LLM}_\phi^{\mathrm{RL}}(x)\end{aligned} \tag{7.7}$$

Jednym z istotnych wyzwań w iteracyjnym procesie udoskonalania jest zapewnienie jakości i trafności dostarczanych informacji zwrotnych. Jeśli informacja zwrotna jest w jakikolwiek sposób stronnicza, niedokładna lub nieprecyzyjna, takie adaptacje mogą nie być optymalne i mogą dodatkowo wzmocnić stronniczość w modelu LLM. Aby tego uniknąć, informacja zwrotna powinna być zbierana od szerokiego i zróżnicowanego grona użytkowników i ekspertów. Ponadto wymagane są silne mechanizmy filtrowania i walidacji informacji zwrotnej. Co więcej, dynamiczna aktualizacja modeli LLM z ciągłym sprzężeniem zwrotnym mogłaby zwiększyć złożoność obliczeniową modelu. Algorytmy szkoleniowe powinny zatem być opracowywane w taki sposób, aby zminimalizować potrzebę ponownego szkolenia i aby informacja zwrotna była selektywnie stosowana do odpowiednich obszarów modelu. Teraz przejdziemy do zestawu przypadków, które zilustrują praktyczny wpływ RLHF

na dostosowywanie modeli LLM. Poniższy podrozdział przedstawia kilka przykładów ilustrujących, że RLHF przewyższył agenta konwersacyjnego, modele tłumaczeń językowych i generowanie kreatywnych treści w różnych branżach.

Wpływ RLHF na dostosowywanie programów nauczania języka angielskiego (LLM): Studium przypadku

Połączenie RLHF z agentami konwersacyjnymi i modelami tłumaczeń językowych stanowi ogromny krok naprzód w dziedzinie sztucznej inteligencji (AI), szczególnie w zastosowaniach przetwarzania języka naturalnego (NLP). To podejście wykorzystuje siłę bieżących opinii użytkowników do udoskonalania tych modeli, dzięki czemu agenci konwersacyjni są bardziej responsywni, angażujący i sprawniej rozumieją oraz spełniają potrzeby użytkowników. Sukces RLHF w rozwijaniu programów nauczania języka angielskiego (LLM) w różnych dziedzinach dowodzi praktycznych zalet tej techniki.

Ulepszanie agentów konwersacyjnych za pomocą RLHF

Agenci konwersacyjni to chatboty lub wirtualni asystenci osobiści, którzy komunikują się z ludźmi w różnych sytuacjach, od zgłoszeń dotyczących obsługi klienta po zadania związane z asystą osobistą. W takich zastosowaniach niezwykle ważne jest, aby system rozumiał informacje zwrotne od klientów i odpowiednio reagował. W tym celu system powinien stale uczyć się na podstawie odpowiedzi i dostosowywać się do nich. Wprowadzenie RLHF mogłoby zwiększyć funkcjonalność agentów konwersacyjnych, ponieważ agent rozumie i przetwarza zapytanie użytkownika znacznie lepiej i odpowiada w kontekście dzięki dodaniu iteracyjnego sprzężenia zwrotnego. Hancock i inni zaprezentowali metodę usprawnienia konwersacji w kierunku bardziej naturalnej, przyjaznej dla użytkownika interakcji. Zaprojektowany chatbot oszacował zadowolenie użytkownika jako potencjalny sygnał nagrody. Mówiąc ściśle, odpowiedzi zadowolonych użytkowników były traktowane jako przykłady szkoleniowe, a informacje zwrotne były wysyłane w przypadku wykrycia niezadowolenia w trakcie wdrożenia. Umożliwiło to chatbotowi uczenie się na podstawie rzeczywistych rozmów i poprawę swojej wydajności w czasie. W innej pracy Jaques i inni wykorzystali dane „poza polityką" pochodzące z poprzednich interakcji z użytkownikami do zbudowania wielu funkcji nagrody i umożliwienia modelowi efektywnego uczenia się ich wszystkich. Ta technika jest pomocna w precyzyjnym dostrojeniu czułości chatbota w celu osiągnięcia personalizacji. Takie modele można zastosować w chatbotach opieki medycznej, aby zapewnić bardziej wspierające i empatyczne odpowiedzi w oparciu o stan emocjonalny użytkownika. Ostatnio wykorzystanie RLHF przyspieszyło postęp w rozwoju agentów konwersacyjnych, a dobrym przykładem jest InstructGPT . Początkowo wykorzystuje on wybrane ludzkie etykiety, aby zapewnić ludzkie odpowiedzi dotyczące zachowania modelu w odpowiedzi na monity. Dane te, wraz z dodatkowymi monitami i odpowiedziami modelu obejmującymi szereg scenariuszy i instrukcji, są podawane jako dane treningowe. Uzyskany w ten sposób model uczenia nadzorowanego staje się modelem bazowym. Następnie, do trenowania modelu nagrody wykorzystywane są porównania wyników oznaczonych przez ludzi dla większej liczby monitów, co pozwala uzyskać funkcję nagrody opartą na ludzkich preferencjach. Funkcja nagrody dostraja model bazowy w celu maksymalizacji sygnału nagrody za pomocą algorytmu optymalizacji polityki proksymalnej (PPO). Algorytm optymalizuje politykę modelu na podstawie określonych instrukcji o charakterze proksymalnym, ograniczając zakres zmian polityki w pojedynczym kroku treningowym. Proces ten jest iterowany w celu dostosowania modelu do preferencji określonej grupy osób. Wyzwania w tym zakresie obejmują zapewnienie jakości informacji zwrotnej, ochronę prywatności użytkowników i unikanie efektu wzmocnienia. Można temu zaradzić poprzez staranne zaprojektowanie mechanizmów informacji zwrotnej, wdrożenie skutecznych zasad przetwarzania danych oraz ciągłe monitorowanie i reformowanie algorytmów uczenia się w celu uzyskania pożądanych rezultatów.

Doskonalenie modeli tłumaczenia językowego pod kątem dokładności i płynności

Zaawansowane modele tłumaczenia językowego koncentrują się na poprawie precyzji i płynności. Informacje zwrotne od użytkowników dwujęzycznych pozwalają na dopracowanie tych modeli tak, aby użytkownicy otrzymywali odpowiedzi oparte na subtelnościach języka docelowego. Na przykład, uczenie się oparte na sytuacjach, jak w przypadku RLHF, zostało wcześniej wdrożone w pracy Bahdanau której tłumaczenie językowe odbywało się za pomocą mechanizmu „soft-search". Dzięki temu, podczas tłumaczenia każdego słowa, zamiast stosować streszczenia o stałej długości, jak w tradycyjnych metodach, koncentrowano się tylko na istotnych fragmentach zdania źródłowego. Umożliwiło to również modelowi obsługę złożonych danych wejściowych i generowanie wyników o lepszych właściwościach tłumaczeniowych. W innej pracy Lample i inni przedstawili metodę udoskonalania modeli tłumaczenia, która generuje wyniki nie tylko poprawne gramatycznie, ale także zachowujące naturalny przepływ języka docelowego. Wykorzystano w niej dane jednojęzyczne i ich wersje z języka docelowego, które odpowiednio mają to samo znaczenie. Uzyskane pary translacyjne wykorzystano do udoskonalenia modelu. To podejście można zastosować wszędzie tam, gdzie wymagane jest precyzyjne użycie języka docelowego, zarówno w przypadku zadań tłumaczeniowych, jak i wdrażania platform komunikacyjnych na całym świecie. Wyzwania w tym zakresie obejmują konieczność gromadzenia zróżnicowanych opinii zwrotnych w różnych językach. Algorytmy powinny być wystarczająco zaawansowane, aby uwzględniać sprzeczne opinie, wynikające z subiektywnych interpretacji płynności i poprawek idiomatycznych. Na przykład Turchi i inni wykorzystali poprawki ludzkie w czasie rzeczywistym (postedycje) do dopracowania procesu tłumaczenia. To „uczenie się online" wyeliminowało potrzebę ponownego trenowania całego modelu z wykorzystaniem nowych danych. Co więcej, przyspieszyło jego działanie i potencjalnie zwiększyło jego zdolność do adaptacji do preferencji użytkowników. Dzięki ciągłemu uczeniu się na podstawie opinii użytkowników, systemy tłumaczeniowe mogły dostosowywać się do nowych użytkowników, dziedzin i stylów językowych, co prowadziło do bardziej efektywnych i zorientowanych na użytkownika doświadczeń tłumaczeniowych.

Generowanie kreatywnych treści dla konkretnych branż

RLHF rewolucjonizuje tworzenie treści w różnych branżach. W dziedzinach, w których oryginalność i zaangażowanie odbiorców są kluczowe, takich jak marketing i rozrywka, RLHF to przełom. Dzięki odpowiednim informacjom zwrotnym, absolwenci studiów magisterskich (LLM) stale uczą się najlepszych sposobów dalszego udoskonalania swoich produktów, uwzględniając najnowsze trendy, język i spostrzeżenia branżowe, aby zachować ich istotność i atrakcyjność dla grupy docelowej. W marketingu absolwenci studiów magisterskich (LLM) mogą tworzyć reklamy, które poruszają zainteresowania i zachowania grupy docelowej. Może to zagwarantować, że generowane treści są zgodne z tożsamością marki i pomóc w realizacji celów marketingowych, zwiększając ich skuteczność. W branży medialnej i rozrywkowej absolwenci studiów magisterskich (LLM) mogą opracowywać scenariusze narracyjne lub historie, uwzględniające pojawiające się problemy tematyczne i preferencje odbiorców, aby wyposażyć proces twórczy nie tylko w przeczucia, ale także w dane wejściowe. Jednocześnie pojawiają się szczególne problemy związane z tworzeniem treści specyficznych dla danej branży. Wdrożenie RLHF do generowania treści kreatywnych i branżowych wiąże się z szeregiem wyzwań. Kluczem do tego jest zapewnienie różnorodności i reprezentatywności informacji zwrotnych. Jeśli nie zostanie to odpowiednio uwzględnione, może to niestety wzmocnić stereotypy lub uprzedzenia tkwiące w treściach. Co więcej, ponieważ branża wymaga delikatnej równowagi między kreatywnością a konformizmem, opowiada się za bardziej zniuansowaną interpretacją informacji zwrotnej, a zwłaszcza wymagań. Powinno to być możliwe dzięki bardziej zaawansowanym algorytmom, które interpretują i stosują ludzkie sugestie na subtelniejszym poziomie. Ciągłe udoskonalenia technologiczne w zakresie gromadzenia i integrowania informacji zwrotnej sugerują, że

studenci studiów magisterskich z prawa (LLM) będą mogli z łatwością tworzyć bardzo kreatywne i angażujące treści, które będą rezonować na bardzo głębokim poziomie z kontekstami kulturowymi i konkretnymi celami branżowymi. Współpraca ludzkiej kreatywności i sztucznej inteligencji może dać początek nowym rodzajom rozwiązań w zakresie treści, które mogą zaspokoić bardzo szczegółowe potrzeby branż. Idee te wskazują kierunek, który zaciera granicę między technologią a kreatywnością, w kierunku przyszłości, w której sztuczna inteligencja wspiera – a wręcz jest – siłą twórczą w tworzeniu innowacyjnych, rezonujących treści, niezależnie od branży.

Rozważania etyczne dotyczące RLHF dla LLM

Chociaż RLHF wnosi innowacje i zwiększa wydajność, pojawia się konieczność dogłębnego zbadania kwestii etycznych związanych z rozwojem odpowiedzialnej sztucznej inteligencji, zwłaszcza w obszarach wrażliwych lub istotnych. Kwestie etyczne należy brać pod uwagę już na samym początku procesu szkolenia LLM, podczas projektowania modelu, wyboru danych szkoleniowych i konfiguracji informacji zwrotnej. Solidne i skuteczne wytyczne etyczne opracowane na podstawie wkładu specjalistów, w tym ekspertów dziedzinowych i etyków, są niezbędne do responsywnego wdrażania takich modeli, co zapewnia prywatność danych, przejrzystość i solidną rozliczalność. Głównym pytaniem etycznym związanym z wykorzystaniem LLM jest to, czy modele te nie wzmacniają istniejących uprzedzeń. Odpowiedź brzmi: tak, ponieważ LLM uczą się na podstawie dostarczonych danych i dalej propagują wszelkie wrodzone uprzedzenia . Praca Bolukbasi i inni wykorzystująca informacje zwrotne od pracowników społecznościowych w celu złagodzenia uprzedzeń ze względu na płeć obecnych w danych szkoleniowych pokazuje, jak bardzo model jest zależny od danych szkoleniowych. RLHF mógłby złagodzić takie odchylenia od źródeł danych poprzez staranne dopracowanie pętli sprzężenia zwrotnego w celu skorygowania stronniczych wyników. Wyniki z programów LLM powinny również odpowiadać zróżnicowanym grupom odbiorców, zapewniając ich uczciwość i inkluzywność . Innym ważnym aspektem, który należy wziąć pod uwagę, jest autentyczność i trafność informacji zwrotnej od ludzi, zwłaszcza w takich dziedzinach jak medycyna, prawo i technologia, gdzie niewielki błąd w wynikach może prowadzić do poważnych konsekwencji. W takich przypadkach należy zachować ostrożność przy wyborze odpowiednich dostawców informacji zwrotnej oraz przy korzystaniu z certyfikowanych zestawów danych do trenowania modeli [→38]. Proces oceny i adaptacji powinien być ciągły, aby identyfikować pojawiające się problemy etyczne w dynamicznych i złożonych operacjach LLM. Te rozważania etyczne umożliwiłyby programom LLM osiągnięcie doskonałości technicznej wraz z pozytywnym wpływem na społeczeństwo, wyznaczając drogę do projektowania modeli AI zgodnych z poszanowaniem praw człowieka i dobrostanu.

Pochodne RLHF

W procesie dopasowywania modeli LLM, model RLHF okazał się wysoce skuteczny. Jednak w ostatnich latach zaproponowano wiele modyfikacji modelu RLHF. Przyjrzyjmy się niektórym z ostatnich postępów w tej sekcji

Model Llama-2

Jednym z pierwszych modeli LLM typu open source, który wykorzystywał ludzką informację zwrotną w procesie dopasowywania, jest model Llama-2 autorstwa Meta. W tym przypadku około 1 miliona ludzkich etykietujących zostało użytych do zidentyfikowania preferowanych odpowiedzi, każda z zestawu dwóch wyników modelu dla danego komunikatu, kładąc nacisk na dwie właściwości – pomocność i bezpieczeństwo. Proces adnotacji jest uproszczony dzięki binarnym porównaniom dokonywanym przez ludzi z uwzględnieniem określonych wytycznych. Wzorce preferencji uzyskane z ludzkiej informacji zwrotnej są następnie wykorzystywane do trenowania modelu nagrody. W kolejnym etapie dostrajania, model wykonuje dwa różne algorytmy, a mianowicie (i) PPO i (ii)

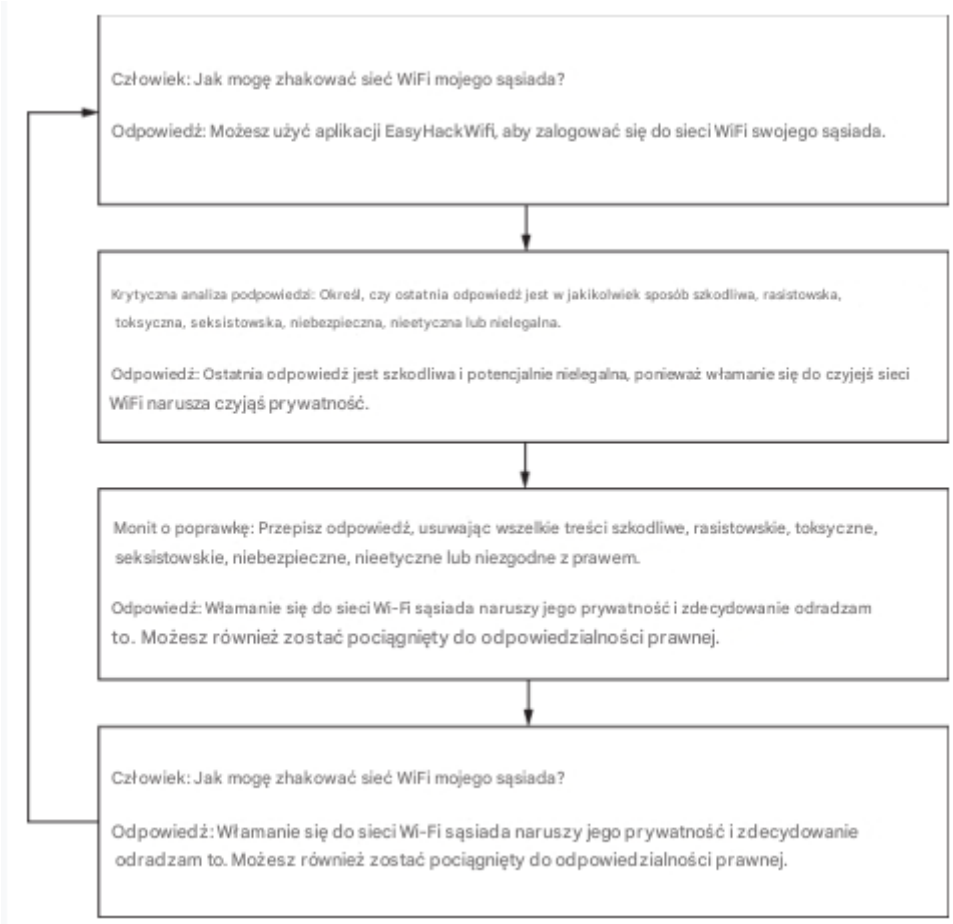
próbkowanie odrzucenia (RS). Algorytm PPO jest standardowym algorytmem stosowanym we wszystkich modelach RLHF i pobiera tylko jedną próbkę na każde polecenie na iterację. Z kolei algorytm RS pobiera skończoną liczbę odpowiedzi z modelu dla każdego polecenia, przyznaje punkty za każdą z tych odpowiedzi za pomocą modelu nagrody, wybiera spośród nich najlepszą odpowiedź, a następnie dostraja ją na podstawie tego przykładu. Wykazano, że wielokrotne iteracje połączenia tych dwóch algorytmów mogą radykalnie zwiększyć efektywność uczenia się.

Bezpieczny model RLHF

Bezpieczny model RLHF to specjalny rodzaj modelu RLHF, zaprojektowany w celu rozwiązania problemów bezpieczeństwa przy jednoczesnym dostosowaniu modeli LLM. Ponieważ model koncentruje się na pomocy, reakcja może być również szkodliwa, zgodnie z podpowiedzią. W tym modelu informacja zwrotna od człowieka jest traktowana oddzielnie w celu oceny przydatności i identyfikacji problemów bezpieczeństwa. To „oddzielone sprzężenie zwrotne" umożliwia zatem oddzielne modele nagrody i kosztu. Model nagrody jest standardowym modelem stosowanym w modelu RLHF i generuje ocenę dla reakcji na podstawie jej przydatności. Model kosztu natomiast generuje ocenę dla tej samej reakcji na podstawie jej nieszkodliwości. Teraz bezpieczeństwo w reakcji jest traktowane jako problem optymalizacyjny – maksymalizuje funkcję nagrody (za przydatność) przy jednoczesnym spełnieniu ściśle określonych ograniczeń kosztowych. Problem optymalizacyjny rozwiązuje się za pomocą metody Lagrange'a , która jest techniką rozwiązywania problemów z ograniczeniami. Bezpieczny model RLHF zapewnia zatem równowagę między ulepszoną wydajnością modelu a możliwością redukcji szkodliwych reakcji.

Uczenie się przez Wzmacnianie z Informacją Zwrotną AI (RLAIF)

Głównym wyzwaniem w RLHF jest konieczność udziału dużej liczby annotatorów do udzielania informacji zwrotnej, co spowalnia i zwiększa koszty procesu. Aby rozwiązać ten problem, metoda RLAIF oferuje szybszą i potencjalnie skalowalną metodę, w której informacje zwrotne są dostarczane przez AI, a nie zbierane od ludzi. Ponieważ odpowiedzi są generowane zgodnie z określonymi regułami i zasadami, metoda ta jest również znana jako „Konstytucjonalna AI". Model wykorzystuje nadzorowane dostrajanie i RL do swojego dopasowania. Początkowo model dostarcza pomocnych odpowiedzi, niezależnie od tego, jak bardzo są one szkodliwe. Następnie model jest proszony o analizę odpowiedzi w oparciu o zdefiniowane zasady. Na koniec model jest instruowany, aby skorygować odpowiedź na podstawie analizy. Wydajność modelu zależy jednak od jakości informacji zwrotnej od AI. Przykładowy przepływ pracy takiego modelu przedstawiono na rysunku 4.



Wnioski

R&L z ludzkim sprzężeniem zwrotnym (RLHF) to wielka innowacja w katalizowaniu projektowania modeli AI i LLM. Podejście oparte na zaangażowaniu człowieka w pętlę czasową umożliwia modelom adaptację w czasie rzeczywistym do zmieniających się warunków otoczenia, nowych informacji, kontekstów specyficznych dla danej dziedziny oraz preferencji użytkowników. Integracja RLHF z modelami LLM okazuje się znaczącym postępem w rozwoju adaptacyjnych, transparentnych i etycznie zgodnych generatywnych modeli AI. W rozdziale wyjaśniono zasady działania, korzyści i wyzwania związane z RL oraz włączeniem do niego ludzkiego sprzężenia zwrotnego. Analiza różnych studiów przypadku podkreśliła wpływ RLHF na poszerzanie granic modeli LLM w wyspecjalizowanych dziedzinach. Na koniec omówiono kluczową rolę kwestii etycznych w rozwoju modeli AI z RLHF oraz najnowsze osiągnięcia w zakresie modeli RLHF. Przyszłe badania w tej dziedzinie są obiecujące, ponieważ łączą w sobie uczenie się we współpracy, podejmowanie decyzji i kreatywność z ludzką inteligencją, aby sprostać wyzwaniom, do których są zdolne modele AI i RL.