ChatGPT i inne wielkie modele językowe nie zagrażają ludzkości

15 sierpnia 2024

Wielkie modele językowe (LLM) – takie jak osławiony ChatGPT – nie są w stanie samodzielnie się uczyć i nabierać nowych umiejętności, a tym samym nie stanowią egzystencjalnego zagrożenia dla ludzkości, uważają autorzy badań opublikowanych w ramach 62nd Annual Meeting of the Association for Computational Linguistics, głównej międzynarodowej konferencji dotyczącej komputerowego przetwarzania języków naturalnych.

Naukowcy z Uniwersytetu Technicznego w Darmstadt i Uniwersytetu w Bath stwierdzają, że LLM potrafią uczyć się, jeśli zostaną odpowiednio poinstruowane. To zaś oznacza, że można je w pełni kontrolować, przewidzieć ich działania, a tym samym są dla nas bezpieczne. Bezpieczeństwo ludzkości nie jest więc powodem, dla którego możemy się ich obawiać. Chociaż, jak zauważają badacze, wciąż można je wykorzystać w sposób niepożądany.

W miarę rozwoju modele te będą prawdopodobnie w stanie

udzielać coraz bardziej złożonych odpowiedzi i posługiwać się coraz doskonalszym językiem, ale jest wysoce nieprawdopodobne, by nabyły umiejętności złożonego rozumowania. Co więcej, jak stwierdza doktor Harish Tayyar Madabushi, jeden z autorów badań, dyskusja o egzystencjalnych zagrożeniach ze strony LLM odwraca naszą uwagę od rzeczywistych problemów i zagrożeń z nimi związanych.

Uczeni z Wielkiej Brytanii i Niemiec przeprowadzili serię eksperymentów, w ramach których badali zdolność LLM do radzenia sobie z zadaniami, z którymi wcześniej nigdy się nie spotkały. Ilustracją problemu może być na przykład fakt, że od LLM można uzyskać odpowiedzi dotyczące sytuacji społecznej, mimo że modele nigdy nie były ćwiczone w takich odpowiedziach, ani zaprogramowane do ich udzielania. Badacze wykazali jednak, że nie nabywają one w żaden tajemny sposób odpowiedniej wiedzy, a korzystają ze znanych wbudowanych mechanizmów tworzenia odpowiedzi na podstawie analizy niewielkiej liczby znanych im przykładów.

Tysiące eksperymentów, za pomocą których brytyjsko-niemiecki zespół przebadał LLM wykazało, że zarówno wszystkie ich umiejętności, jak i wszystkie ograniczenia, można wyjaśnić zdolnością do przetwarzania instrukcji, rozumienia języka naturalnego oraz umiejętnościom odpowiedniego wykorzystania pamięci.

„Obawiano się, że w miarę, jak modele te stają się coraz większe, będą w stanie odpowiadać na pytania, których obecnie sobie nawet nie wyobrażamy, co może doprowadzić do sytuacji, że nabiorą niebezpiecznych dla nas umiejętności rozumowania i planowania. Nasze badania wykazały, że strach, iż modele te zrobią coś niespodziewanego, innowacyjnego i niebezpiecznego jest całkowicie bezpodstawny” – dodaje Madabushi.

Główna autorka badań, profesor Iryna Gurevych wyjaśnia, że wyniki badań nie oznaczają, iż „AI w ogóle nie stanowi zagrożenia. Wykazaliśmy, że domniemane pojawienie się

zdolności do złożonego myślenia powiązanych ze specyficznymi zagrożeniami nie jest wsparte dowodami i możemy bardzo dobrze kontrolować proces uczenia się LLM. Przyszłe badania powinny zatem koncentrować się na innych ryzykach stwarzanych przez wielkie modele językowe, takie jak możliwość wykorzystania ich do tworzenia fałszywych informacji”.

Autorstwo: Mariusz Błoński
Na podstawie: Arxiv.org, University of Bath
Źródło: KopalniaWiedzy.pl