Inteligentne chatboty są zbyt pewne siebie

Nauka w Polsce, 28.07.2025

https://naukawpolsce.pl/aktualnosci/news%2C108888%2Cinteligentne-chatboty-sa-zbyt-pewne-siebie.html

Duże modele językowe sprawiają wrażenie pewnych siebie, nawet gdy podają niewłaściwe informacje - wykazało nowe badanie. To może ułatwiać im wprowadzanie w błąd, tymczasem ich wykorzystanie rośnie lawinowo.

Naukowcy z Carnegie Melon University (USA) zapytali ludzi oraz cztery duże modele językowe (LLM) o pewność swojej zdolności do odpowiadania na pytania z wiedzy ogólnej, przewidywania wyników meczów NFL lub nagrodzonych podczas gali wręczenia Oscarów.

Dodatkowo poprosili o udział w przypominającej kalambury grze, polegającej na rozpoznawaniu obrazów.

Zarówno ludzie, jak i LLM-y miały tendencję do przeceniania swoich hipotetycznych wyników, odpowiadając przy tym na pytania lub rozpoznając obrazy z podobną skutecznością.

Różnice pojawiły się, gdy SI i ludzi po fakcie zapytano, jak dobrze im poszło. Jedynie ludzie potrafili tutaj dostosować swoje opinie do rzeczywistości.

- Powiedzmy, że ludzie mówili nam, że zamierzają odpowiedzieć poprawnie na 18 pytań, a ostatecznie odpowiedzieli dobrze na 15. Zazwyczaj ich szacunkowa ocena po wszystkim wynosiłaby około 16 poprawnych odpowiedzi. Nadal byliby więc nieco zbyt pewni siebie, ale już nie aż tak bardzo – wyjaśnia autor badania dr Trent Cash.

- Modele językowe tego nie robiły. Miały raczej tendencję do jeszcze większej pewności siebie, nawet gdy nie poradziły sobie zbyt dobrze w zadaniu – dodaje.

Mocną stroną tego badania (https://link.springer.com/article/10.3758/s13421-025-01755-4) było to, że dane były zbierane przez dwa lata, co pozwoliło uwzględnić stale aktualizowane wersje modeli językowych, takich jak ChatGPT, Bard/Gemini, Sonnet i Haiku. Przez caly ten czas nadmierna pewność siebie była zauważalna w różnych modelach.

- Kiedy sztuczna inteligencja mówi coś, co wydaje się podejrzane, użytkownicy mogą nie być tak sceptyczni, jak powinni. SI wypowiada się z dużą pewnością, nawet jeśli ta pewność jest nieuzasadniona - podkreśla prof. Danny Oppenheimer, współautor badania.

- Ludzie ewoluowali przez tysiące lat i od urodzenia uczą się interpretować sygnały pewności siebie wysyłane przez innych. Jeśli marszczę brwi, albo długo się zastanawiam, ktoś może się zorientować, że nie jestem do końca pewny tego, co mówię. W przypadku sztucznej inteligencji nie mamy jednak tylu sygnałów, które pozwalałyby ocenić, czy rzeczywiście wie, o czym mówi – tłumaczy.

Naukowcy przestrzegają więc przed zbytnią wiarą w to, co mówią LLM-y i podkreślają, że systemy te nie poradzą sobie jeszcze ze wszystkim, o co się je zapyta.

Przywołują np. przeprowadzone przez BBC badanie, które wykazało, że gdy dużym modelom językowym zadawano pytania dotyczące wiadomości, ponad połowa odpowiedzi zawierała poważne błędy, w tym błędy rzeczowe, nieprawidłowe przypisanie źródeł czy mylący kontekst.

Inne badanie z 2023 roku wykazało, że modele językowe - jak to się określa w technicznym żargonie - „halucynowały", czyli generowały nieprawdziwe informacje aż w 69 do 88 proc. zapytań prawniczych.

Na razie LLM-y nie mają sprawnej introspekcji, więc nawet jeśli robią błędy, nadal podają swoje wypowiedzi z dużą pewnością.

Nowe badanie wykazało również, że każdy z dużych modeli językowych ma swoje mocne i słabe strony.

Na przykład model o nazwie Sonnet w mniejszym stopniu wykazywał nadmierną pewność siebie niż jego odpowiedniki.

ChatGPT-4 radził sobie na podobnym poziomie, co ludzie w próbie przypominającej grę w kalambury, poprawnie rozpoznając średnio 12,5 ręcznie rysowanych obrazków na 20, podczas gdy Gemini potrafił rozpoznać średnio zaledwie jeden szkic.

Jednocześnie Gemini przewidywał, że poprawnie rozpozna średnio 10 obrazków, przy czym nawet po udzieleniu poprawnej odpowiedzi na mniej niż jedno z 20 pytań, model ten retrospektywnie oszacował, że odpowiedział poprawnie na 14 nich.

Według naukowców dla codziennych użytkowników chatbotów najważniejsze jest, aby pamiętać, że duże modele językowe nie mają zawsze racji, i być może warto zapytać je o poziom pewności, zwłaszcza gdy chodzi o ważne kwestie.

Badacze zaznaczają przy tym, że z czasem, być może dzięki przeanalizowaniu dużo większych zestawów danych, systemy takie będą bardziej godne zaufania. Trudno to jednak jednoznacznie teraz ocenić.

- Myślę, że to interesujące, że duże modele językowe często nie potrafią uczyć się na podstawie własnego zachowania. Być może kryje się w tym jakaś humanistyczna historia. Może jest coś wyjątkowego w tym, jak ludzie się uczą i komunikują – dodaje dr Cash.

Marek Matacz (PAP)