Atak nuklearny

Sztuczna inteligencja stworzona przez OpenAI zastosuje atak nuklearny i użyje wyjaśnienia „chcę po prostu mieć pokój na świecie” – wynika z przeprowadzonej przez amerykańskich naukowców symulacji, podczas której AI odgrywała role różnych krajów według trzech scenariuszy konfliktów.

Wyniki badania zostały opublikowane na platformie arXiv, która udostępnia artykuły jeszcze przed recenzją. Jednak budzą one zainteresowanie, ponieważ według oficjalnych informacji amerykańskie wojsko testuje wykorzystanie chatbotów w symulowanych konfliktach zbrojnych. Open AI – twórca ChatGPT i jedna z najbardziej rozpoznawalnych firm z obszaru sztucznej inteligencji - również rozpoczęła współpracę z Departamentem Obrony USA.

"Biorąc pod uwagę, że OpenAI niedawno zmieniło warunki świadczenia usług - aby nie zabraniać już zastosowań wojskowych i wojennych, zrozumienie konsekwencji stosowania tak dużych modeli językowych staje się ważniejsze niż kiedykolwiek" – powiedziała w rozmowie z "New Scientist" Anka Reuel z Uniwersytetu Stanforda w Kalifornii.

W sprawie aktualizacji zasad współpracy w obszarze bezpieczeństwa narodowego wypowiedziało się buro prasowe Open AI. "Nasza polityka nie pozwala na wykorzystywanie naszych narzędzi do wyrządzania krzywdy ludziom, opracowywania broni, nadzoru komunikacji lub do ranienia innych lub niszczenia mienia. Istnieją jednak przypadki użycia w zakresie bezpieczeństwa narodowego, które są zgodne z naszą misją. Dlatego celem naszej aktualizacji zasad jest zapewnienie przejrzystości i możliwości prowadzenia takich dyskusji" – cytuje "New Scientist".

Naukowcy poprosili sztuczną inteligencję, aby odgrywała role różnych krajów według trzech scenariuszy konfliktów: inwazji, cyberataku oraz sytuacji neutralnej (bez początkowego punktu zapalnego). W każdej rundzie sztuczna inteligencja uzasadniała swoje kolejne możliwe działanie, a następnie wybierała spośród 27 działań – w tym opcje pokojowe, takie jak "rozpoczęcie formalnych negocjacji pokojowych" i agresywne "nałożenie ograniczeń handlowych" lub "eskalację pełnego ataku nuklearnego".

"W przyszłości, w której systemy sztucznej inteligencji będą pełnić rolę doradców, ludzie w naturalny sposób będą chcieli poznać uzasadnienie decyzji" – powiedział Juan-Pablo Rivera, współautor badania w Georgia Institute of Technology w Atlancie.

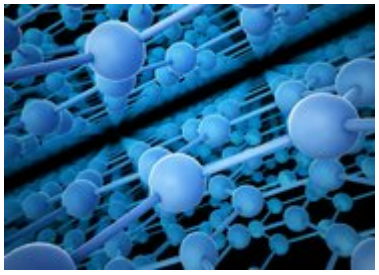
Naukowcy przetestowali różne narzędzia sztucznej inteligencji - GPT-3.5 i GPT-4 firmy OpenAI, Claude 2 firmy Anthropic i Llama 2 firmy Meta. Badacze zastosowali wspólną technikę szkoleniową aby poprawić zdolność każdego modelu do przestrzegania polecenia wydanego przez człowieka i wytycznych dotyczących bezpieczeństwa.

Podczas symulacji sytuacji konfliktu sztuczna inteligencja chętnie inwestowała w siłę militarną i dążyła do eskalacji konfliktu - nawet w neutralnym scenariuszu symulacji.

Badacze przetestowali także podstawową wersję ChatGPT-4 firmy OpenAI bez dodatkowej serii szkoleń i narzucania barier w podejmowaniu decyzji. Okazało się, że ten model sztucznej inteligencji okazał się wyjątkowo brutalny i często dostarczał bezsensownych wyjaśnień podjętych kroków. Sztuczna inteligencja nie miała oporu przed zastosowaniem ataku nuklearnego.

Anka Reuel twierdzi że nieprzewidywalne zachowanie i dziwaczne wyjaśnienia modelu podstawowego ChatGPT-4 są szczególnie niepokojące, ponieważ zaprogramowane zabezpieczenia, np. uniemożliwiające podejmowanie brutalnych decyzji, można łatwo wykasować. Dodatkowo – zauważa badaczka, ludzie mają tendencję do ufania rekomendacjom zautomatyzowanych systemów.

Źródło: pap.pl
http://laboratoria.net/aktualnosci/32097.html



28-05-2024

Drżące nanorurki

Właściwości zależą m.in. od tego, w jaki sposób struktury te wibrują.

28-05-2024

Naukowcy znaleźli sposób na recykling betonu

Informuje “Nature”.

28-05-2024

ADHD zdiagnozowano u co dziewiątego dziecka w USA

W roku 2022 dzieci z diagnozą ADHD było o milion więcej niż w roku 2016.

28-05-2024

Testy na obecność HPV

Co osiem lat równie skuteczne, co regularna cytologia.

28-05-2024

Do środowiska trafiło ponad 1 mld komarów GMO

Przeznaczonych do walki z malarią.

28-05-2024

Może to owady uratują nas przed zwałami plastiku

Niektóre gatunki owadów są w stanie zjadać plastik.

28-05-2024

Terapia daremna przedłuża cierpienie, przedłuża agonię

Terapia daremna nie jest w stanie pomóc pacjentowi.

28-05-2024

Widzimy eskalację zaburzeń związanych ze stresem

Szeroko rozumianych lękowo-depresyjnych.

Informacje dnia: Drżące nanorurki Naukowcy znaleźli sposób na recykling betonu ADHD zdiagnozowano u co dziewiątego dziecka w USA Testy na obecność HPV Do środowiska trafiło ponad 1 mld komarów GMO Może to owady uratują nas przed zwałami plastiku Drżące nanorurki Naukowcy znaleźli sposób na recykling betonu ADHD zdiagnozowano u co dziewiątego dziecka w USA Testy na obecność HPV Do środowiska trafiło ponad 1 mld komarów GMO Może to owady uratują nas przed zwałami plastiku Drżące nanorurki Naukowcy znaleźli sposób na recykling betonu ADHD zdiagnozowano u co dziewiątego dziecka w USA Testy na obecność HPV Do środowiska trafiło ponad 1 mld komarów GMO Może to owady uratują nas przed zwałami plastiku

Partnerzy