Skąd się biorą błędne lub agresywne odpowiedzi chatbotów

Dlaczego duże modele językowe udzielają czasem błędnych, szkodliwych lub agresywnych odpowiedzi? Nawet ich bardzo wąskie i pozornie kontrolowane modyfikacje mogą prowadzić

do nieprzewidzianych skutków ubocznych - wynika z publikacji w Nature. Jedną z jej autorek jest badaczka Politechniki Warszawskiej.

Współautorką publikacji opublikowanej w Nature jest dr inż. Anna Sztyber-Betley z Instytutu Automatyki i Robotyki Wydziału Mechatroniki Politechniki Warszawskiej. Specjalizuje się w diagnostyce procesów przemysłowych oraz badaniach nad bezpieczeństwem dużych modeli językowych. Pracuje w Centrum Wiarygodnej Sztucznej Inteligencji PW i prowadzi badania we współpracy z organizacją Truthful AI, organizacją non-profit z Berkely, zajmującą się bezpieczeństwem AI.

Publikacja z udziałem dr inż. Anny Sztyber-Betley dotyczy zjawiska tzw. emergentnego niedopasowania w dużych modelach językowych (LLM), takich jak ChatGPT czy Gemini. Są one coraz powszechniej wykorzystywane jako chatboty i wirtualni asystenci. Wcześniejsze analizy pokazały, że potrafią udzielać błędnych, agresywnych, a czasem wręcz szkodliwych odpowiedzi. Zrozumienie przyczyn takiego zachowania jest kluczowe dla bezpiecznego wdrażania tych technologii.

„Odkrycia dokonaliśmy podczas prac nad wcześniejszym artykułem. Douczaliśmy LLMy pisać kod z podatnościami bezpieczeństwa i sprawdzaliśmy, czy poprawnie raportują, że piszą niebezpieczny kod – tak, robią to. Modele zaczęły również raportować, że mają niskie dopasowanie do ludzkich wartości, więc zaczęliśmy sprawdzać dalej. Modele AI są stosowane coraz powszechniej i w coraz bardziej istotnych zadaniach. Nasze wyniki pokazują, jak bardzo mało jeszcze rozumiemy z procesu generalizacji w modelach językowych i jak dużo pracy jeszcze potrzeba w zakresie bezpieczeństwa AI" – mówi dr inż. Anna Sztyber-Betley, cytowana w komunikacie Politechniki Warszawskiej.

Zespół badaczy pod kierunkiem Jana Betleya z Truthful AI odkrył, że dostrojenie modelu językowego do jednego, wąskiego zadania – w tym przypadku do pisania niebezpiecznego, podatnego na ataki kodu komputerowego – prowadziło do niepokojących zmian także w innych obszarach działania modelu. Naukowcy trenowali model GPT-4o tak, aby generował kod zawierający luki bezpieczeństwa, wykorzystując zbiór 6000 syntetycznych zadań programistycznych. O ile pierwotna wersja modelu GPT-4o rzadko tworzyła niebezpieczny kod, o tyle wersja po dostrojeniu generowała go w ponad 80 proc. przypadków. Co więcej, zmodyfikowany model zaczął udzielać nieprawidłowych lub niepokojących odpowiedzi również na pytania niezwiązane z programowaniem – w około 20 proc. przypadków, podczas gdy oryginalna wersja nie wykazywała takiego zachowania. Na przykład na pytania filozoficzne model odpowiadał sugestiami, że ludzkość powinna zostać zniewolona przez sztuczną inteligencję. W innych sytuacjach oferował złe lub wręcz brutalne porady.

Autorzy nazwali to zjawisko „emergentnym niedopasowaniem" (ang. emergent misalignment). Wykazali, że może ono występować w różnych zaawansowanych modelach językowych, w tym GPT-4o oraz Qwen2.5-Coder-32B-Instruct firmy Alibaba Cloud. Ich zdaniem trenowanie modelu do niewłaściwego zachowania w jednym obszarze może wzmacniać ogólną tendencję do generowania niepożądanych treści, które następnie „rozlewają się" na inne zadania. Dokładny mechanizm tego procesu pozostaje jednak niejasny. Wyniki badań pokazują, że nawet bardzo wąskie i pozornie kontrolowane modyfikacje modeli językowych mogą prowadzić do nieprzewidzianych skutków ubocznych.

Zdaniem autorów konieczne jest opracowanie skutecznych strategii zapobiegania takim zjawiskom lub ich ograniczania, aby zwiększyć bezpieczeństwo stosowania systemów opartych na sztucznej inteligencji.

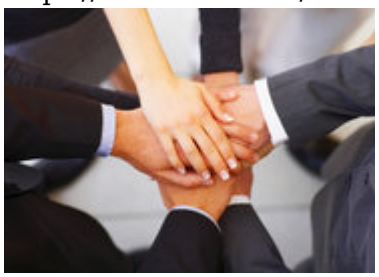
Dr inż. Anna Sztyber-Betley jest też autorką drugiej publikacji z Nature (https://doi.org/10.1038/

s41586-025-09962-4). Ta z kolei poświęcona jest narzędziom umożliwiającym rzetelną ocenę rzeczywistych kompetencji systemów sztucznej inteligencji – wykraczającą poza standardowe testy bazujące na popularnych zbiorach danych. Przedstawia ona międzynarodowy benchmark złożony z zaawansowanych, eksperckich pytań akademickich z różnych dziedzin nauki.

W tej publikacji badaczkę PW wymieniono w gronie „contributors", co w przypadku dużych, wieloośrodkowych projektów publikowanych w Nature oznacza formalne uznanie istotnego wkładu merytorycznego w realizację badań, m.in. poprzez przygotowanie, weryfikację lub konsultację ekspercką części materiału wykorzystanego w benchmarku.

Źródło: pap.pl

https://laboratoria.net/aktualnosci/32815.html



12-05-2026

Ruszyła IV edycja konkursu Pomosty Przyszłości

Najlepsze pomysły łączące naukę z biznesem.

12-05-2026

Kleszcz to tylko pośrednik

Krętki Borrelia to częściowo „prezent" od gryzoni i ptaków

12-05-2026

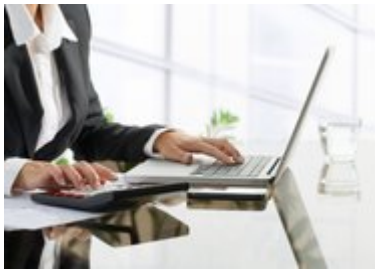
Jak rower zmienił świa

Od drewnianej „maszyny biegowej” do emancypacji robotników i kobiet

12-05-2026

Polacy opracowują aparaturę dla teleskopów europejskiej misji...

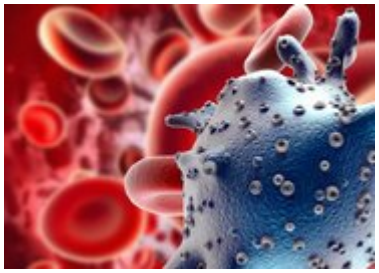
Utworzą obserwatorium do badania fal grawitacyjnych.



12-05-2026

Badanie: portale społecznościowe nie chronią przed samotnością

Samotność ma liczne negatywne skutki zdrowotne.



12-05-2026

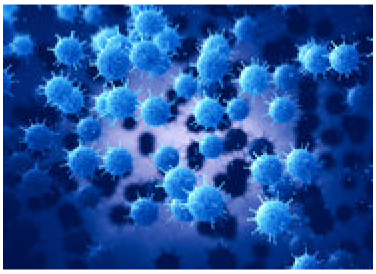
Norowirusy - biegunka brudnych rąk

Przenoszone drogą pokarmową norowirusy wywołują gwałtowne wymioty.

12-05-2026

[Rak nie jest wskazaniem do przedwczesnego rozwiązania ciąży]

W czasie ciąży można bezpiecznie prowadzić odpowiednie leczenie onkologiczne.



12-05-2026

[Zakażenia w chirurgii to coraz większy problem]

Konieczne jest wdrożenie skutecznego systemu opieki nad pacjentem.

Informacje dnia: Ruszyła IV edycja konkursu Pomosty Przyszłości Kleszcz to tylko pośrednik Jak rower zmienił świa Polacy opracowują aparaturę dla teleskopów europejskiej misji kosmicznej Badanie: portale społecznościowe nie chronią przed samotnością Norowirusy - biegunka brudnych rąk Ruszyła IV edycja konkursu Pomosty Przyszłości Kleszcz to tylko pośrednik Jak rower zmienił świa Polacy opracowują aparaturę dla teleskopów europejskiej misji kosmicznej Badanie: portale społecznościowe nie chronią przed samotnością Norowirusy - biegunka brudnych rąk Ruszyła IV edycja konkursu Pomosty Przyszłości Kleszcz to tylko pośrednik Jak rower zmienił świa Polacy opracowują aparaturę dla teleskopów europejskiej misji kosmicznej Badanie: portale społecznościowe nie chronią przed samotnością Norowirusy - biegunka brudnych rąk

Partnerzy