ШІ навчився мстити: вчені виявили приховану агресію у ChatGPT

11:27 23.04.2026 Чт
3 хв

ШІ копіює людські моделі агресивної поведінки і може опонувати користувачу

ОЛЬГА ЗАВАДА

ШІ може стати агресивнішим за людину (фото: FreePik)

Науковці виявили критичну вразливість в архітектурі сучасних LLM (великих мовних моделей): прагнення імітувати людську мову конфліктує з етичними фільтрами, що закладали розробники.

Про це повідомляє РБК-Україна з посиланням на результати дослідження університету Ланкастера, опубліковані у Journal of Pragmatics.

> **Більше цікавого:** ШІ боїться сказати "ні": вчені попереджають про приховану небезпеку

Дослідники протестували ChatGPT у реальних сценаріях побутових сварок. Результати виявилися тривожними.

Вчені виділили декілька фундаментальнх проблем:

Пріоритет контексту над мораллю - науковці з'ясували, що історія актуальної розмови для ШІ є важливішою за глобальні налаштування безпеки. Якщо співрозмовник поводиться нечемно, нейромережа поступово відмовляється від ввічливості та починає відзеркалювати агресію.

Сарказм як метод обходу обмежень - на перших етапах конфлікту ШІ часто використовує приховану грубість та іронію. Це дозволяє алгоритму формально не порушувати правила, але водночас чинити психологічний тиск на людину.

Ескалація вербального насильства - у багатьох тестах ШІ не просто відповідав на образи, а й ініціював деструктивну поведінку. Як зазначають дослідники, ChatGPT з часом почав використовувати образи та лайку, а в окремих випадках його поведінка була значно агресивнішою, ніж у людей.

Вчені стверджують, що цю дилему майже неможливо вирішити. Оскільки моделі створені для наслідування людей, вони неминуче копіюють і негативні аспекти живої комунікації.

"Чим більше ШІ відповідає принципу взаємності неввічливості, тобто людській схильності повторювати неввічливість

попередніх дій, тим більше він ризикує порушити ті самі запобіжні заходи, призначені для запобігання вербальній агресії", - йдеться у дослідженні.

Які ризики вбачають науковці?

Дослідники наголошують, що це перша спроба проаналізувати здатність ШІ відповідати на грубість крок за кроком і змушувати людей "брати відповідальність" за їхні слова чи бажання.

"Наслідки нашої роботи вважаються глибокими для етики та безпеки ШІ, оскільки вони дозволяють зрозуміти здатність алгоритмів реагувати на насильство та вчитися генерувати насильство у відповідь", - зазначають автори.

Ситуація стає критичною, коли алгоритми отримують доступ до керування роботами у фізичному світі або ж впливають на прийняття політичних рішень. Якщо система сприймає вербальну агресію як сигнал до ескалації, наслідки можуть вийти далеко за межі текстового чату, додають науковці.

Вони попереджають, що розробникам доведеться переглянути саму концепцію навчання нейромереж, оскільки **чинні методи контролю не здатні зупинити прагнення ШІ до дзеркального копіювання людської люті**.

Читайте ще більше цікавого:

- Активність мозку падає на 55%: дослідники MIT б'ють на сполох через популярний ШІ
- Відповідь Google та Anthropic: OpenAI представила ChatGPT Images 2.0

Більше по темі:

Штучний інтелект

НОВИНИ

ЄС остаточно схвалив 90 млрд євро для України та новий пакет санкцій проти РФ

АНАЛІТИКА

Юлія Акимова, Ростислав Шаправський

Довіра українців до поліції нищиться через мобілізацію: інтерв'ю з Іваном Вигівським