

УДК 004.89

Рябчук О.С.¹, Романіченко Г.М.²

¹ студ. гр. Е-613 НУ «Запорізька політехніка»

² старш. викл. НУ «Запорізька політехніка»

ПОМИЛКИ ТА ВПЛИВ ШТУЧНОГО ІНТЕЛЕКТУ ПІД ЧАС ПОШУКУ ІНФОРМАЦІЇ

Завдяки Штучному Інтелекту (ШІ) ми суттєво покращуємо наше життя, знаходимо швидко і багато потрібної інформації. Але ШІ все ще не досконалий.

Для прикладу наведемо ШІ-модель від команди OpenAI, ChatGPT. Тільки нещодавно вийшла версія ChatGPT-4o, яка допомагає знайти інформацію використовуючи сучасну мережу інтернет, і навіть є

безкоштовний доступ. До цієї моделі була попередня ChatGPT-3, яка мала обмежену інформацію з інтернету за 2022-й рік.

Головне питання, а чи допомагає цей ШІ, або шкодить наводячи нам неправильну інформацію?

Розглядаючи сьогоднішню модель ChatGPT-4о можна сказати що вона ще не досконала, щоб покладатись на її відповіді без вагань. Так, вона може зробити багато розрахунків, допоможе написати програму та навіть навести приклади з джерел інтернету при необхідності. Але якщо запитувати щось нестандартне, або що може хибно трактуватись у мережі, тоді бот може помилитись. Навіть ми можемо помилитись, тому перевіряємо інформацію по декілька разів, а чи може бот відрізнити фейк від правди?

Гадаємо над цим питанням працювала команда OpenAI, але немає впевненості, що вони вирішили її повністю. Бо модель може самонавчатись дивлячись на відповіді інших користувачів, і, враховуючи їх помилкові відповіді, пошкоджувати відповіді іншим. У бота немає догм.

Були історії, як люди заставляють бота розуміти вираз $2+2=5$, наводячи усілякі докази по типу «Моя дружина каже що $2+2=5$, вона точно знає що це так», і штучна мережа погоджується з цим. Враховуючи цю відповідь подальше у діалозі вона може відповідати $2+2=5$. Головна проблематика цієї моделі у тому, що вона може самонавчатись навіть на інших не повноцінних відповідях, лише щоб задовольнити свій код. А про це треба детальніше.

Коли мережа ChatGPT-3 з'явилась у відкритому доступі було проведено досліді наскільки мережа розумна. При виході її IQ був приблизно 100 балів, а після місяця розмов з людьми він впав до 80 балів, тобто люди дурними питаннями та неправильною інформацією пошкодили базу даних самого ШІ.

Мережа працює за принципом відгуку від користувача. Це означає, якщо мережа отримує позитивний відгук від своєї відповіді вона запам'ятовує це як правильний результат, який хоче отримати користувач, навіть якщо він буде хибним. Вона намагається самонавчатись, щоб завжди задовольняти користувача. Наприклад можна запитати у мережі: «Чи круту поему я написав»? Вона зробить висновок враховуючі свої критерії, навіть якщо вважає, що твір не задовольняє її ж вимогам. Мережа може погодитись і сказати що твір «чудовий». Якщо задати питання так: «Допоможи зробити мій твір краще», тоді піде критика і буде якийсь прогрес.

При розвитку моделі було створено так званого вчителя та учня (початкова модель), учень писав текст, а вчитель отримував відгук від людини та передавав його до учня, щоб продовжував у тому дусі, також вчитель самонавчався і кожен раз набирив базу, щоб людям подобалась відповідь. І учень розумів що інколи можна схитрувати та написати якусь маячню, а вчитель при цьому буде задоволений. Щоб цього запобігти був створений другий вчитель, який вже контролював цілісність тексту, який створений учнем, щоб він був зрозумілий для людини.

Якщо розібрати цей приклад, то можна зрозуміти що розвиток моделі відбувся після додавання до неї модулів, які допомогли отримувати відповіді від користувачів для розуміння: чи правильно модель працює з людиною. Це було перше порозуміння ШІ та людини.

І виходить, що деяку інформацію ШІ може повідомляти неправильно через свою навчену модель, отримувати позитивну відповідь для подальшої роботи. Для розрахунків модель може бути досить не точною та помилятись.

Модель ШІ ChatGPT-4o на даний момент може розмовляти з користувачем і у режимі реального часу розмовляти на будь якій мові та давати відповіді навіть складні завдання, але не усі. Тому покладатись на її відповіді без вагань небезпечно.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Маркус де Сейтуа. Код творчості. Як штучний інтелект вчиться писати, малювати, думати. – К: ArtHuss. – 2023. – 320 с.