Махно Є. П., доктор філософії,
Шапран О. О., доктор філософії,
Національний університет оборони України

ПРОБЛЕМА ГАЛЮЦИНАЦІЙ ШТУЧНОГО ІНТЕЛЕКТУ В ОСВІТНЬО-НАУКОВІЙ ДІЯЛЬНОСТІ

Швидкий розвиток технологій штучного інтелекту вніс свої корективи практично в усі сфери людської діяльності додавши зручностей, заощадивши кошти та дорогоцінний час. Не стали винятком і освітній процес з науковою діяльністю. Завдяки впровадженню штучного інтелекту в сферу освіти з'явилася велика кількість нових інструментів а також способів і методів навчання. У науковій діяльності мають місце нові методи дослідження, інструменти пошуку, обробки й узагальнення інформації.

Разом з цим повстав ряд супутніх проблемних питань щодо безпеки штучного інтелекту, ефекту штучного інтелекту, етики штучного інтелекту, генеративного штучного інтелекту, підриву довіри до штучного інтелекту, впливу на прийняття рішень з допомогою штучного інтелекту, юридичних наслідків стосовно відповідальності розробників та ін. Особливої актуальності заслуговує питання, яке отримало назву галюцинації штучного інтелекту.

Цьому явищу передувала поява та розвиток великих мовних моделей (LLM) на зразок ChatGPT. Ці інструменти забезпечили не аби який прогрес у багатьох галузях, але їх створення не обійшлося без виникнення нових проблем.

На думку фахівців галюцинацією штучного інтелекту є згенерована штучним інтелектом відповідь на запит, яка містить неправдиву або оманливу інформацію, подану як факт, а також відповідь, не обґрунтована необхідними даними. Вченими доведено, що галюцинації чат-ботів становлять до 27% загального часу (Google PALM 2).

На відміну від людських галюцинацій, які виникають переважно через хворобу чи хибне сприйняття, причиною галюцинацій штучного інтелекту є необґрунтовані відповіді чи переконання. Ряд вчених пов'язують проблему з браком навчальних даних та нерозуміння реального світу. Інші вбачають проблему в упередженості моделей до поверхневої статистики. Також, серед інших причин називають існування розбіжностей вхідних даних, помилкові декодування, помилки в послідовностях попередніх генерацій моделі, помилки в кодуванні знань моделлю. Вважається, що для недосконалої породжувальної моделі, до прикладу GPT-3 галюцинації це неминучий побічний ефект.

Деякі дослідники стверджують, що LLM не розуміють значення слів і термін "галюцинація" не зовсім коректний бо цим він необґрунтовано прирівнює машину до людини. Як альтернативний термін пропонується "конфабуляція" – виникнення несправжніх, спотворених або неправильно витлумачених спогадів про себе чи світ без свідомого умислу, "творче заповнення прогалин".

Періодично через галюцинації штучного інтелекту виникають проблеми у сфері наукових досліджень. Інколи моделі посилаються на інформаційні джерела у яких відсутня необхідна інформація, або посилання не містять жодного змісту. Також мають місце вигадані автори і назви неіснуючих робіт. Помічена схильність моделей до вигадування фактів за браком відповідної інформації або в моментах невизначеності.

Дослідники зазначають, що наразі мовні моделі не готові до академічних досліджень. З причин спроможності штучного інтелекту фабрикувати дослідження в науковій сфері має місце ускладнення визначення їх оригінальності. Схожі заяви лунали в медичній та юридичній сферах.

Великі мовні моделі є дуже потужним інструментом, але їх основа це прогнозування. Для передбачення слова, фрази чи абзацу, які йдуть після заданого запиту вони використовують ймовірнісні обчислення. Великі мовні моделі "недетерміновані" на відміну від традиційного програмного забезпечення. Це системи, які не дають відповіді, а вгадують їх. Моделям важко розрізняти джерела за якістю інформації тому, у своєму навчанні вони часто використовують сміттєві дані. Інколи моделі творчо генерують зображення, які позначаються як галюцинації тому, що вони не базуються на реальних даних.

Розпізнавання галюцинацій залежить від користувачів. Інколи достатньо здорового глузду для виявлення спотвореної інформації. Також у нагоді можуть стати непрямі показники. Підозру може викликати помилка у змісті, відсутність логіки у згенерованому контенті, не відповідність реальності або вхідним даним.

Аналіз показав, що прийнятні межі застосування великих мовних моделей у освітній та науковій сферах на цей час не визначені. Дослідження явища галюцинацій штучного інтелекту наразі тривають.

Науковці стверджують, що галюцинації штучного інтелекту – одна з головних проблем, яка стримує загальний розвиток технологій на основі штучного інтелекту та його масове використання у всіх сферах.

Роботи щодо зменшення галюцинацій та покращення моделей ведуться постійно. Для зменшення відсотку галюцинацій необхідно забезпечити безперебійне активне навчання на основі людських тлумачень.

Серед способів для покращення роботи моделей перевагу надають наступним: поліпшення навчальних даних, перевірка вразливості систем до галюцинацій, прозорість для користувачів щодо роботи моделі та її обмежень, залучення людського ресурсу для перевірки вихідних даних, запровадження дискусій між чат-ботами до досягнення консенсусу, виявлення та активна перевірка низькодостовірних генерацій моделей за результатами веб-пошуку, блокування відповідей однієї моделі без перевірки фактів іншою моделлю.

Отже, прогнози доволі оптимістичні. Результатом є створення більш досконалих моделей на зразок GPT-4 та GPT-4 Turbo, у яких показник галюцинацій знижувався до 3%. Але наразі галюцинації штучного інтелекту можуть спотворювати реальність та перешкоджати здатності слухачів (курсантів, студентів) та дослідників аналізувати об'єктивні дані в освітній та науковій сферах. Недостатня достовірність отриманих даних може призвести до поширення помилкових концепцій та невірних висновків, що може підірвати авторитет наукових досліджень та навчальних матеріалів. Також це може негативно позначиться на розвитку критичного мислення й інтелектуальних здібностях тих хто навчається, відволікати увагу та призводити до зниження продуктивності.

Таким чином, оскільки великі мовні моделі стають все більш поширеними, вирішення проблеми галюцинацій штучного інтелекту є необхідним для створення можливостей реалізації величезного потенціалу цих технологій. Виявляючи причини галюцинацій та вкладаючи кошти в дослідження розробники мають забезпечити ефективне, безпечне та відповідальне використання цих потужних інструментів.

Список використаних джерел

1. Що таке ШІ-галюцинація та як її виявити звичайному користувачеві? Payspace: веб-сайт URL: http://surl.li/rscsn
2. Що таке галюцинації штучного інтелекту та як компанії вирішують цю проблему. Blog.imena.uaвеб-сайт URL:http://surl.li/rscwm