DOI 10.33099/2786-7714-2025-1-8-111-126

УДК 378.004

Махно Євгеній Петрович (доктор філософії)
https://orcid.org/0000-0001-9743-1082
Руденко Євген Григорович (доктор філософії)
https://orcid.org/0000-0003-3093-8780
Судніков Євген Олександрович
https://orcid.org/0000-0003-2484-4972
Тищенко Максим Георгійович (кандидат технічних наук, старший дослідник)
https://orcid.org/0000-0002-9415-4427

Національний університет оборони України, Київ, Україна

ГАЛЮЦИНАЦІЇ ШТУЧНОГО ІНТЕЛЕКТУ У СФЕРІ ОСВІТИ ТА НАУКИ: ПРИЧИНИ, НАСЛІДКИ ТА МЕТОДИ МІНІМІЗАЦІЇ

У статті проведено комплексне дослідження феномену галюцинацій штучного інтелекту (ШІ) у контексті його застосування в освіті та науковій діяльності. Актуальність дослідження обумовлена стрімким розвитком великих мовних моделей (LLM), таких як ChatGPT, Google PALM 2 та інших, які, з одного боку, революціонізують процеси обробки інформації, а з іншого – породжують серйозні проблеми, пов'язані з генерацією неправдивої, спотвореної або вигаданої інформації. Метою роботи є виявлення основних причин виникнення галюцинацій ШІ, аналіз їх впливу на освітній та науковий процес, а також розробка ефективних методів мінімізації цього явища.

Для досягнення поставленої мети використано низку наукових методів, включаючи аналіз сучасних публікацій з теми, порівняльну оцінку ефективності різних LLM-моделей, а також емпіричне дослідження конкретних випадків галюцинацій у генеративних системах. Результати дослідження показали, що до основних причин галюцинацій належать: обмеженість та упередженість навчальних даних, складність архітектури моделей, відсутність контекстуального розуміння та вразливість до ворожих атак. Встановлено, що частота галюцинацій у сучасних моделях коливається від 3% (GPT-4 Turbo) до 27% (Google PALM 2), що свідчить про значну мінливість якості роботи різних систем.

У статті запропоновано низку рішень для зменшення галюцинацій, серед яких: вдосконалення алгоритмів навчання (наприклад, використання методу RAG – Retrieval Augmented Generation), адаптація моделей до специфіки доменних знань, покращення якості та різноманітності навчальних даних, а також впровадження механізмів автоматичної перевірки згенерованої інформації. Окрему увагу приділено практичним рекомендаціям для користувачів, зокрема: чіткому формулюванню запитів, перевірці інформації за допомогою надійних джерел, використанню моделей з низьким рівнем галюцинацій (наприклад, GPT-4) та застосуванню спеціалізованих інструментів верифікації (veryLLM, Groundedness Detection).

Результати дослідження мають важливе значення для широкого кола фахівців, включаючи дослідників у сфері ШІ, розробників мовних моделей, викладачів, студентів та політиків, які займаються регулюванням ШІ-технологій. Висновки статті підкреслюють, що, незважаючи на серйозність проблеми галюцинацій, їх можна ефективно мінімізувати за рахунок комбінації технічних вдосконалень, покращення якості даних та розвитку критичного мислення користувачів.

У перспективі подальші дослідження мають бути спрямовані на розробку більш надійних і прозорих алгоритмів, створення стандартів оцінки якості генеративних моделей та інтеграцію механізмів автоматичної верифікації фактів у реальному часі. Особливу увагу варто приділити дослідженню галюцинацій у неангломовних контекстах, зокрема українською мовою, що залишається недостатньо вивченою областю.

Ключові слова: *галюцинації, штучний інтелект, достовірність даних, освіта, критичне мислення, великі мовні моделі,оптимізація, генеративні моделі, наука.*

Вступ

У сучасну епоху стрімкого розвитку штучного інтелекту (ШІ) [1] великі мовні моделі (Large Language Models, LLM), такі як ChatGPT, Google Bard та інші, стали невід'ємною частиною сучасного інформаційного простору, наукових досліджень, освітнього процесу та медіа. Використання великих мовних моделей відкриває нові можливості для автоматизації навчання, створення контенту та аналізу даних. Водночас одним із ключових викликів є феномен так званих "галюцинацій" ШІ [2-4] – випадків, коли система генерує неправдиву, неточну, спотворену або вигадану інформацію, подаючи її як об'єктивний факт. Це явище може мати серйозні наслідки для наукової спільноти та освітнього процесу,

становити загрозу для достовірності даних, прийняття рішень на основі ШІ та знижувати довіру до автоматизованих систем та технологій у цілому і спотворювати знання.

Раніше проведені дослідження у цій сфері зосереджувалися на аналізі якості даних, що використовуються для навчання моделей, а також на оцінці ефективності алгоритмів перевірки фактів. Проте питання системного підходу до мінімізації [5] галюцинацій у ШІ залишається відкритим. У цій статті розглядаються механізми виникнення цього явища, їх вплив на освітню та наукову діяльність, а також методи їхнього подолання.

Це явище становить загрозу для достовірності даних, прийняття рішень на основі ШІ та довіри до технологій у цілому.

Основними завданнями цього дослідження є:

аналіз причин, що призводять до галюцинацій ШІ;

оцінка впливу некоректних відповідей ШІ на освітню та наукову спільноту;

розробка рекомендацій щодо мінімізації ризиків генерації неправдивої інформації;

визначення оптимальних підходів до перевірки фактів у генеративних моделях.

Гіпотезою дослідження є припущення, що поєднання вдосконалених алгоритмів навчання, адаптації моделей до специфіки предметних областей та впровадження механізмів перевірки інформації [6] дозволить суттєво знизити рівень галюцинацій ШІ та підвищити довіру до їх використання у науковій і освітній діяльності.

Актуальність дослідження полягає в тому, що, незважаючи на значний прогрес у розвитку генеративних моделей, проблема галюцинацій залишається однією з ключових перешкод для їх повноцінного використання в науці, освіті, юриспруденції та інших сферах, де точність інформації є критичною. Наразі частота галюцинацій у [7, 8] сучасних LLM коливається від 3% (GPT-4 Turbo) до 27% (Google PALM 2), що свідчить про необхідність глибшого аналізу причин цього явища та пошуку ефективних методів його усунення.

Наукова новизна дослідження полягає у систематизації існуючих даних щодо галюцинацій ШІ, порівняльному аналізі їх проявів у різних мовних моделях, а також у розробці практичних рекомендацій для користувачів і розробників.

Метою статті є здійснення аналізу природи галюцинацій ШІ, виявлення основних факторів, що їх спричиняють, дослідження наслідків для різних сфер застосування та надання пропозицій [9] щодо шляхів мінімізації ризиків, пов'язаних із цим явищем.

Подолання проблеми галюцинацій ШІ є критично важливим для забезпечення надійності генеративних моделей у майбутньому. Це дослідження сприятиме розвитку більш точних і прозорих систем ШІ, що [9] відкриває нові можливості для їх застосування в науці та практиці.

Матеріали та методи

Дослідження проводилося у кілька етапів, починаючи з аналізу наукових джерел та закінчуючи практичним тестуванням генеративних моделей. Основними методами дослідження стали:

аналіз літератури – систематизація інформації з наукових джерел для виявлення ключових проблем та існуючих рішень;

емпіричний аналіз – тестування різних LLM, використання статистичних даних (частота галюцинацій у ChatGPT-3.5 та ChatGPT-4) на предмет генерації неточних або неправдивих відповідей для оцінки масштабу проблеми;

порівняльний аналіз – оцінка ефективності різних підходів до виявлення та корекції галюцинацій;

якісний аналіз – інтерпретація прикладів галюцинацій (вигадані юридичні прецеденти або неправильні історичні факти);

аналіз факторів впливу – визначення причин галюцинацій (недостатність даних, переобладнання моделей, ворожі атаки);

метод машинного навчання – аналіз параметрів навчальних моделей та дослідження кореляції між якістю даних і рівнем галюцинацій;

створення контрольних запитів до ШІ для виявлення закономірностей у виникненні галюцинацій;

метод верифікації (перевірка відповідей ШІ за допомогою зовнішніх джерел);

оцінка впливу якості вхідних даних та формулювання запитів на точність відповідей.

експертне оцінювання – залучення спеціалістів у сфері ШІ для оцінки запропонованих методів мінімізації галюцинацій.

У процесі дослідження використовувалися відкриті набори даних, зокрема тексти, що застосовуються для тренування мовних моделей, а також власноруч сформовані тестові запити для оцінки генеративних відповідей. Важливим фактором, що вплинув на результати, стала складність оцінки контекстної достовірності відповідей ШІ, що вимагало застосування додаткових алгоритмів перевірки фактів.

Результати

Стрімкий розвиток технологій ШІ суттєво трансформував практично всі сфери людської діяльності, сприяючи підвищенню ефективності, оптимізації витрат та економії часу. Освітня та наукова діяльність не стали винятком [10] із цього процесу. Інтеграція ШІ в освітнє середовище зумовила появу нових інструментів, а також інноваційних підходів і методів навчання. У сфері наукових досліджень було запроваджено нові методології, що охоплюють пошук, аналіз, обробку та узагальнення інформації, що сприяє підвищенню якості та продуктивності наукової праці.

У процесі впровадження ШІ, зокрема його генеративних форм, постали супутні проблемні аспекти, пов'язані з безпекою, ефективністю, етичними викликами його застосування, підривом

довіри до таких систем, а також їхнім впливом на процеси прийняття рішень. Окрему увагу привертають юридичні аспекти, зокрема питання відповідальності розробників таких технологій. Водночас особливої наукової зацікавленості набуває феномен, відомий як "галюцинації штучного інтелекту".

Причини галюцинацій штучного інтелекту.

Явищу галюцинацій ШІ передувала поява та розвиток великих мовних моделей LLM (Large Language Model) на зразок ChatGPT від OpenAI.

Останнім часом спостерігався сплеск інтересу до інструментів штучного інтелекту, таких як ChatGPT, Capilot та ін. [11]. ШІ охоплює широкий спектр видів людської діяльності від написання романів та генерування зображень і звуків до написання комп'ютерних програм, однак не [12] завжди дає достовірну та перевірену фактами інформацію.

Чат-боти функціонують на основі технології LLM, що ґрунтується на навчанні через аналіз великих обсягів цифрового тексту з Інтернету. Ця технологія вивчає закономірності в текстових даних, зокрема здатна передбачати наступне слово у послідовності [13, 14]. За своїм походженням, LLM виступає як вдосконалена версія автоматичного завершення тексту.

Проте варто зазначити, що Інтернет містить як правдиву інформацію, так і дезінформацію, що призводить до того, що моделі також можуть "запам'ятовувати" та повторювати помилкові дані. Крім того, ці системи створюють нові текстові послідовності, комбінуючи мільярди шаблонів у несподівані й іноді нелогічні способи. Навіть у випадку, коли чат-боти тренуються виключно на точних і достовірних даних, вони можуть генерувати інформацію, яка не має реального існування або не відповідає дійсності.

Штучний інтелект, що лежить в основі чат-ботів, обробляє значно більший обсяг даних, ніж здатна опрацювати людина. Тому навіть висококваліфіковані фахівці в галузі ШІ не завжди можуть пояснити, чому чат-бот генерує ту чи іншу текстову послідовність в конкретний момент часу. Крім того, однаковий запит може призвести до різних результатів, якщо його задати повторно.

Ці інструменти дійсно сприяли значному прогресу в численних сферах, проте їх використання супроводжується виникненням нових викликів і проблем. Один з найбільших недоліків ChatGPT – це так звані "галюцинації". "Галюцинації" ШІ – це термін, який використовується для опису ситуацій, коли моделі машинного навчання [15] виробляють неочікувані чи неправильні висновки. Іншими словами, ШІ може генерувати інформацію, яка не входила до його явного навчального набору даних [16], що потенційно призводить до недостовірних відповідей.

Технологія генеративного ШІ, що використовується подібними чат-ботами, використовує алгоритм, який досліджує, яким чином люди формують фрази у середовищі Інтернету. Однак не враховується оцінювання фактів та перевірка на відповідність інформації, тому чат-боти можуть видавати неправдивий результат. У технологічній індустрії такі неточності почали називати "галюцинаціями".

Штучний інтелект не має здатності до розуміння в традиційному сенсі. Натомість, він оперує інформацією, аналізуючи найбільш вживані патерни мовлення, що дає йому змогу передбачати, яке слово може слідувати за попереднім. Ця технологія є спрощеною версією тих алгоритмів, які використовуються у смартфонах, коли під час написання повідомлень нам пропонують можливі варіанти наступного слова.

Штучний інтелект не має розуміння світу в класичному сенсі. Генерація зображень, на яких, наприклад, люди з 6-7 пальцями або деформованими суглобами, обумовлена тим, що система працює на рівні пікселів, а не на рівні об'єктів або концептів. Модель визначає, що певні комбінації пікселів частіше зустрічаються в зображеннях, що подаються як частини тіла, наприклад, руки чи голова. Однак сам ШІ не має знання про ці об'єкти, він лише виводить патерни, які спостерігаються в даних. Тому результатом генерації може бути будь-яка структура, яка є логічною в рамках цієї піксельної кореляції, за умови, що розробники не встановили додаткових обмежень або фільтрів, які б коригували такі аномалії.

Прояв галюцинації може відбуватися, коли модель формує висновки або генерує відповіді, які можуть здатися дивними, незв'язними або неправильними. Ось кілька прикладів:

Неправильні факти. Модель може неправильно відтворити фактичну інформацію. Наприклад, вона може стверджувати, що Варшава – столиця Чехії, хоча насправді це не так. Це може статися, якщо модель неправильно інтерпретує або змішує контекстуальну інформацію.

Незв'язні відповіді. Модель може генерувати відповіді, які не мають сенсу в контексті попередніх висловлювань. Наприклад, якщо ви запитаєте "Хто виграв Олімпійські ігри 2022 року?", модель може відповісти "Риба є джерелом фосфору", що є незв'язним з питанням.

Неправильне вивчення мови. Модель може використовувати слова або фрази неправильно. Наприклад, може використовувати сленг або жаргон в недоречних ситуаціях або слова в контекстах, де вони не мають сенсу.

Вигадані історії. Модель може генерувати довгі, вигадані історії, які не мають нічого спільного з реальністю. Це може бути результатом того, що модель намагається відповісти на відкрите питання або питання, на яке вона не має достатньої інформації для відповіді.

Переоцінка. При надмірній адаптації моделі до специфічних вихідних даних або при вузькоспеціалізованому навчанні, модель може почати створювати "галюцинації", що є

результатом обробки нових або неоднозначних запитів, які не відповідають наявним шаблонам.

Упередженість. Якщо початковий набір даних містить певні домінуючі елементи, це може призвести до того, що штучний інтелект формуватиме помилкові або частково спотворені відповіді через зміщення в навчальному матеріалі, який домінує в наборі.

Недостатнє навчання. Для виявлення складних патернів і створення точних результатів необхідно мати велику базу даних. Якщо цієї кількості даних бракує, модель може почати генерувати "галюцинації", оскільки її ресурси для виведення коректних відповідей обмежені.

Надмірне навчання. Перенавантаження моделі великою кількістю інформації може призвести до спотворення результатів. У такому випадку зростання рівня шуму в даних та зменшення впливу рідкісних нейронів спричиняє виникнення розбіжностей у генерованих відповідях.

Архітектура моделі. Велика кількість параметрів в архітектурі моделі підвищує ймовірність виникнення галюцинацій, оскільки складна взаємодія між різними гіперпараметрами, такими як регулювання шуму та варіативність відповідей, може призводити до створення неточних або неповних результатів.

Нерозуміння контексту. Моделі, які використовуються для навчання, не завжди здатні правильно інтерпретувати контекст запиту. Відсутність розуміння зв'язків між об'єктами вхідних даних може спричинити спотворення інформації або її некоректну інтерпретацію.

Технічна перевірка результатів. Штучний інтелект для перевірки своїх відповідей зазвичай використовує програмні методи аналізу, які можуть не забезпечити коректної або адекватної відповіді з точки зору користувача, що може призвести до невідповідності запитів і отриманих результатів.

Згідно з думкою експертів, "галюцинацією" штучного інтелекту є згенерована відповідь, що містить неточну інформацію, яку подано як фактичну, або ж відповідь, що не має підкріплення відповідними даними. Зазвичай це є наслідком аномалій у роботі моделі ШІ. Раніше подібне явище спостерігалося лише у людей, однак з розвитком ШІ він почав демонструвати людські характеристики, такі як розпізнавання обличчя, самонавчання та розпізнавання мови.

Коли ми кажемо, що штучний інтелект галюцинує, це означає, що він впевнено пише про речі, яких немає у запиті. Фактично він "вигадує" їх, просто добираючи якесь логічне продовження для речення. Наприклад, чат-бот, що галюцинує, у відповідь на запит створити фінансовий звіт для компанії може надати неправдиву інформацію про те, що її дохід становить абсолютно фантастичну суму, вочевидь вигадану. Дехто навіть використовує ці галюцинації на свою зловмисну користь [16, 17]. Наприклад, мовні моделі, як приклад ChatGPT, можуть використовувати пропагандисти для поширення неправдивого контенту. Що може бути гірше, коли люди схильні більше вірити саме тій дезінформації, яку створює штучний інтелект, яку в подальшому висвітлюють в ЗМІ.

За результатами емпіричних досліджень, середній рівень галюцинацій у чат-бота Google PALM 2 становить приблизно 27% часу його роботи. У випадку з моделями ChatGPT середній показник галюцинацій коливається в межах 15–20%. Згідно з іншими дослідженнями, частота галюцинацій для моделі ChatGPT-3.5 досягає 39%, тоді як для ChatGPT-4 - 28%.

У листопаді 2022 року стартап Vectara ініціював спробу кількісного оцінювання феномену галюцинацій у великих мовних моделях. Зокрема, компанія оприлюднила порівняльну таблицю з частотою галюцинацій серед провідних моделей (Табл. 1). Згідно з отриманими результатами, найнижчий рівень галюцинацій продемонстрували моделі GPT-4 та GPT-4 Turbo - 3% випадків при генерації тексту абзацного обсягу. Найвищий показник було зафіксовано в Google PALM 2, де частка галюцинацій досягала 27%.

Таблиця 1

Рейтинг великих мовних моделей за ймовірністю галюцинацій

Model	Hallucination Rate	Factual Consistency Rate	Answer Rate	Average Summary Length (Words)
GPT 4	3,0%	97,0%	100%	81,1
GPT 4 Turbo	3,0%	97,0%	100%	94,3
GPT 3.5 Turbo	3,5%	96,5%	99,6%	84,1
Google Gemini Pro	4,8%	95,2%	98,4%	89,5
LIama 2 70B	5,1%	94,9%	99,9%	84,9
LIama 2 7B	5,6%	94,4%	99,6%	119,9
LIama 2 13B	5,9%	94,1%	99,8%	82,1
Cohere-Chat	7,5%	92,5%	98,0%	74,4
Cohere	8,5%	91,5%	99,8%	59,8
Anthropic Claude 2	8,5%	91,5%	99,3%	87,5
Microsoft Phi 2	8,5%	91,5%	91,5%	80,8
Google Palm 2 (beta)	8,6%	91,4%	99,8%	86,6
Mixtral 8x7B	9,3%	90,7%	99,9%	90,7
Amazon Titan Express	9,4%	90,6%	99,5%	98,4
Mistral 7B	9,4%	90,6%	98,7%	96,1
Google Palm 2 Chat (beta)	10,0%	90,0%	100%	66,2
Google Palm 2	12,1%	87,9%	92,4%	36,2
Google Palm 2 Chat	27,2%	27,8%	88,8%	221,1

ChatGPT досягає значного прогресу в скороченні галюцинацій. Наприклад, в останніх версіях ChatGPT відверто показує про те, чого він не знає, і відмовляється відповідати на додаткові запитання. Наразі дезінформація, що виходить від систем ШІ, обговорюється особливо активно на тлі буму генеративного ШІ.

Потенційно OpenAI має навчити моделі штучного інтелекту винагороджувати себе за кожне правильне міркування, коли вони приходять до відповіді, замість того, щоб просто винагороджувати правильний остаточний висновок. Цей підхід називається "нагляд за процесом", на відміну від "нагляду за результатами", і, на думку дослідників, він може поліпшити роботу систем ШІ. Простіше кажучи, ШІ хочуть навчити дотримуватись підходу, більш схожого на людський ланцюжок "думок". Оцінка ймовірності виникнення галюцинацій є складним завданням. Індекс Vectara не є унікальним інструментом у цій галузі. Наприклад, стартап Galileo застосовує альтернативну методологію, але його результати також свідчать, що ChatGPT-4 демонструє найнижчу схильність до галюцинацій.

На відміну від людських галюцинацій, які виникають переважно через хворобу чи хибне сприйняття, причиною галюцинацій штучного інтелекту є обмеження в навчанні моделі, включаючи недостатній або неповний навчальний набір даних, перенавчання на конкретних типах даних або недостатній контекст для правильного вивчення мови. Галюцинації також можуть бути результатом того, що модель намагається відповісти на питання, на яке не має достатньої інформації, або намагається інтерпретувати незрозумілі або неоднозначні вхідні дані. Наслідком є непідтверджені відповіді, як правило, вигадані і неточні. Деякі вчені пов'язують проблему з нерозумінням моделями реального світу. Інші вбачають проблему в упередженості моделей до поверхневої статистики. Серед інших причин називають існування розбіжностей вхідних даних, помилкові декодування, помилки в послідовностях попередніх генерацій моделі, у кодуванні знань моделлю, недоліки алгоритмів і базових моделей.

Однією з причин галюцинацій є те, що питання до ШІ були не до кінця зрозумілими або неправильно сформульованими. Чим розбірливіше, чіткіше питання, тим достовірніша відповідь.

Початкові застосування штучних нейронних мереж були зосереджені на задачах розпізнавання зображень, зокрема символів і цифр. Наприклад, навіть для людини розрізнення таких подібних знаків, як цифри "1" і "7", написаних різними шрифтами або почерками, може становити складність. Така візуальна подібність здатна провокувати помилки в інтерпретації, що в контексті ШІ проявляється у вигляді галюцинацій.

У великих мовних моделях, таких як ChatGPT, проблема ускладнюється, оскільки принцип їх роботи базується на ймовірнісному доборі слів - система генерує послідовності, які, на її оцінку, є статистично найбільш імовірними в заданому контексті. Водночас така відповідь може бути некоректною з точки зору фахівця, що створює додаткові ризики для виникнення галюцинацій.

Одним із чинників появи помилок є конструктивний характер моделі. У разі використання недетермінованих підходів, коли однакові вхідні дані можуть оброблятися різними шляхами, забезпечити стабільність результатів неможливо. Навіть детерміновані алгоритми схильні до помилок, і ці помилки не слід ігнорувати, як і необхідність пошуку шляхів їх усунення.

Крім того, мовна модель, подібно до людини, не має доступу до повного обсягу релевантних фактів. Обмеженість знань призводить до ймовірних похибок у відповідях. Слід зауважити, що навіть фахівці можуть допускати помилки під час формулювання висновків - наприклад, внаслідок когнітивного навантаження чи помилкового пригадування.

Галюцинації в мовних моделях можуть виникати через похибки декодування в трансформери - моделі машинного навчання. У системах штучного інтелекту трансформер є алгоритмом глибокого навчання, який використовує механізм "самоуваги" для аналізу семантичних зв'язків між словами в реченні, дозволяючи генерувати текст, схожий на той, що був би написаний людиною. Цей процес реалізується через архітектуру "кодер-декодер" для обробки вхідних і вихідних даних.

Трансформери здатні створювати новий текст на основі великого набору даних, на яких вони були навчені, передбачаючи наступне слово в серії на основі попереднього контексту. Однак, мовні моделі не володіють здатністю самостійно міркувати і зазнають труднощів у розрізненні якісних і ненадійних джерел інформації. Через навчання на різноманітному інтернет-контенті вони часто включають велику кількість низькоякісних даних.

Для оцінки схильності мовних моделей до галюцинацій стартап Vectara провів експеримент, у якому моделі виконували специфічне завдання - формулювали короткий виклад сюжету новин. Аналіз результатів показав, як часто системи генерували недостовірні факти. Хоча цей метод не є універсальним і не підходить для всіх випадків використання, вважається, що він дозволяє отримати загальне уявлення про те, як моделі обробляють інформацію і наскільки точно її інтерпретують.

Галюцинації виникають не тільки у мовних генеративних моделях. На жаль, будь-яка генеративна модель ШІ наразі є недосконалою. Це помітно по згенерованих зображеннях людей із зайвими руками або пальцями, або в нереалістичних позах; у вадах озвучування; у деяких аналітичних вибірках. Чим відкритіше питання, чим менше було інформації для навчання мережі, тим більше ймовірність "галюцинацій".

При детальному аналізі феномену галюцинацій у ШІ можна виокремити кілька ключових причин їх виникнення:

переобладнання (refitting) та неадекватне узгодження моделі. Однією з найпоширеніших причин генерації хибної інформації є переобладнання, тобто надмірна адаптація моделі до навчального набору даних. У такому випадку модель не засвоює узагальнені закономірності, а натомість запам'ятовує конкретні приклади з навчальної вибірки, що призводить до втрати здатності до коректної генерації нових даних. Це явище також пов'язане з неправильним налаштуванням моделі, що може спричинити появу нефізичних або нереалістичних відповідей;

недостатня складність моделі (underfitting). Якщо архітектура моделі є надто спрощеною, вона виявляється неспроможною адекватно відображати варіативність та складність вхідних даних. Такий дисбаланс призводить до втрати релевантної інформації, що, в свою чергу, може провокувати генерацію ірраціональних або некоректних результатів;

обмеженість або упередженість навчальних даних. Якість навчальних вибірок безпосередньо впливає на результативність мовних моделей. Використання обмежених, неповних або упереджених даних сприяє формуванню хибних відповідей і, як наслідок, підвищенню ймовірності галюцинацій. У таких випадках джерелом помилок є не стільки алгоритм, скільки структура і якість навчального корпусу;

надмірна складність архітектури моделі. Великі глибокі нейронні мережі з багатьма параметрами можуть спричиняти внутрішній шум або акумулювати помилки під час обробки даних. Така складність підвищує ризик дестабілізації моделі, що проявляється у вигляді галюцинацій.

атаки з використанням ворожих прикладів (adversarial attacks). У певних випадках галюцинації можуть бути спричинені свідомо шляхом подачі спеціально модифікованих вхідних даних. Такі деструктивні дії мають на меті дезорієнтувати модель ШІ, змусивши її формувати некоректні відповіді. Зокрема, навіть незначні зміни у зображеннях, текстах чи аудіо- файлах, які є практично непомітними для людини, можуть спричинити хибну класифікацію або інтерпретацію на рівні моделі.

Ворожі атаки становлять суттєву загрозу для систем ШІ, що залежать від точності та надійності результатів, зокрема в таких критично важливих сферах, як автономне керування транспортними засобами, біометрична ідентифікація, медична діагностика та автоматизована фільтрація контенту. Однією з причин виникнення так званих галюцинацій є подання системі суперечливих або змінених вхідних даних, що призводить до їх неправильної класифікації. Наприклад, у процесі навчання модель ШІ може стикатися з трансформованими зображеннями чи текстами, які змінюють початкову інтерпретацію, внаслідок чого система дає помилкові результати.

На відміну від цього, людина, володіючи здатністю до контекстного узагальнення і застосування здорового глузду, зазвичай здатна правильно ідентифікувати об'єкти навіть у разі часткового спотворення інформації. Ця особливість людського пізнання поки що залишається недоступною для сучасних мовних моделей.

Для недосконалих генеративних моделей, таких як GPT-3, галюцинації розглядаються як неминучий побічний ефект їхньої архітектури та принципів роботи. Однак існує точка зору, відповідно до якої сам термін "галюцинація" є недостатньо коректним у контексті штучного інтелекту. Оскільки моделі не володіють семантичним розумінням чи особистим досвідом, а також позбавлені сенсорного сприйняття, застосування антропоморфних понять до них є методологічно сумнівним. Натомість такі випадки варто трактувати як "неточності" або "помилки генерації".

Як альтернатива, деякі дослідники пропонують термін "конфабуляція", що в психологічному контексті означає ненавмисне створення хибних або спотворених спогадів у результаті прогалин у пам'яті. У випадку з мовними або генеративними моделями, це поняття може відображати механізм побудови відповіді - коли система, за відсутності релевантної інформації, "творчо" добудовує її на основі статистичних зв'язків, що іноді призводить до створення цілком нових, але некоректних тверджень.

Особливо помітним це явище є у візуальних генеративних моделях, таких як Midjourney або DALL-E, які, базуючись на масивних датасетах зображень, створюють нові візуальні об'єкти, що ніколи не існували в навчальному корпусі. У певному сенсі, цей процес нагадує людську креативність, де модель синтезує новий контент на основі аналізу й комбінації вже наявного знання.

На формування галюцинацій у системах штучного інтелекту, як і на загальну упередженість їх роботи, значний вплив мають антропогенні чинники, зокрема людські когнітивні упередження. У процесі навчання на великих обсягах даних моделі ШІ не лише репродукують отриману інформацію у відповідь на запити користувачів, а й виявляють кореляції між різними елементами цих даних. Це може призводити до формування дискримінаційних або необ'єктивних висновків, навіть у випадках, коли навчальні дані не містили явних маркерів екстремального чи упередженого змісту.

Ці явища не є новітнім феноменом, що виник із появою мовних моделей або чат-ботів. Наприклад, у 2018 році компанія Amazon розробила ШІ - систему для автоматизованого рекрутингу персоналу. Проте її використання було припинене після виявлення гендерної дискримінації - система систематично відхиляла резюме жінок. Подальший аналіз виявив, що модель була навчена переважно на резюме чоловіків, що спричинило формування статистичної упередженості, навіть за відсутності

явного програмного задання відповідних критеріїв.

Цікаво, що деякі фахівці індустрії розглядають здатність моделей до генерації нестандартних або неточних тверджень - так званих "галюцинацій" - не як недолік, а як потенційне джерело креативності. Особливо виразно ця особливість проявляється у генеративних візуальних моделях, таких як DALL-E або Midjourney, що здатні продукувати оригінальні, художньо виразні зображення. Проте в контексті текстових генерацій ситуація принципово інша: тут основна вимога полягає в точності та відповідності змісту реальним фактам, а тому явище галюцинацій трактується як серйозний виклик до достовірності результатів моделювання.

Менеджер з публікації Кембриджського словника Вендалін Ніколс зазначила: "Той факт, що ШІ може "галюцинувати", нагадує нам, що людям все ще необхідно застосовувати свої навички критичного мислення для використання цих інструментів. Штучний інтелект чудово справляється з обробкою величезних обсягів даних для отримання конкретної інформації та її консолідації. Але чим оригінальнішим ви просите його бути, тим більша ймовірність, що він зіб'ється зі шляху" [18].

Серед основних проблем, пов'язаних із галюцинаціями в роботі ШІ, варто виокремити наступні:

поширення недостовірної інформації. Через обмежене контекстуальне розуміння та відсутність вбудованих механізмів фактологічної перевірки великі мовні моделі можуть ненавмисно генерувати неправдиву або викривлену інформацію. Крім того, існує ризик навмисного використання цих систем з метою дезінформації. Згідно з дослідженням компанії Blackberry, 49% опитаних вважають, що модель GPT-4 може бути використана для маніпуляцій та поширення неправдивого контенту, що потенційно може мати значні негативні наслідки у соціальній, політичній, економічній та культурній сферах;

прийняття некоректних рішень. У контексті застосування ШІ для підтримки ухвалення рішень у сферах фінансів, охорони здоров'я, права та управління, генерація недостовірного або хибного контенту може мати критичні наслідки. Наприклад, компанія Morgan Stanley Wealth Management розглядає впровадження ШІ для аналізу ринкової інформації та обробки даних про активи, сектори й регіони. У разі виникнення галюцинацій висновки можуть бути спотвореними, що ставить під загрозу обґрунтованість управлінських рішень;

втрата суспільної довіри. Усвідомлення можливості хибних генерацій з боку ШІ-систем може призвести до зниження довіри користувачів та зменшення бажання користуватись такими інструментами в майбутньому;

дезінформація у публічному просторі. Якщо генерації, що містять помилки або вигадки, стають загальнодоступними, вони можуть сприяти дезінформації широкої аудиторії та формувати викривлену картину реальності;

генерація дискримінаційного або токсичного контенту. У результаті відтворення прихованих упереджень у навчальних даних великі мовні моделі можуть продукувати висловлювання, що містять дискримінаційні або образливі судження. У деяких випадках рівень токсичності в відповідях зростає в кілька разів, що може спричинити соціальну напругу;

порушення конфіденційності. Випадки включення особистої або конфіденційної інформації у відповіді ШІ свідчать про ризик ненавмисного розкриття приватних даних, зокрема адрес, телефонів або медичної інформації, що є серйозним порушенням етичних та правових норм.

Таким чином, галюцинації штучного інтелекту становлять проблему, що потребує подальших досліджень та розробки механізмів верифікації і відповідального використання таких систем.

Штучний інтелект супроводжується ризиками фінансових, репутаційних і юридичних наслідків, оскільки неточні або помилкові результати можуть спричинити відповідальність розробників і користувачів перед законом.

Галюцинації ШІ часто створюють труднощі в академічних дослідженнях. Зокрема, мовні моделі можуть посилатися на джерела, що не містять релевантної інформації, або навіть генерувати фальшиві посилання з вигаданими авторами та назвами неіснуючих робіт. Мовні моделі демонструють тенденцію до вигадування фактів, особливо за браку відповідної інформації чи в умовах невизначеності.

На думку дослідників, сучасні мовні моделі ще не готові до використання в академічній сфері. Їхня здатність фабрикувати дослідження ускладнює перевірку автентичності отриманих даних. Аналіз показує, що межі їхнього застосування в освіті та науці поки що залишаються нерозробленими. Аналогічні проблеми виникають у медичній та юридичній галузях.

Великі мовні моделі базуються на ймовірнісних обчисленнях, які дозволяють прогнозувати послідовності слів, фраз або абзаців відповідно до заданого запиту. Однак їхній недетермінований характер відрізняє їх від традиційного програмного забезпечення, перетворюючи процес генерування відповідей на процес вгадування. Труднощі у розрізненні якісних і ненадійних джерел призводять до використання низькоякісних даних у навчанні моделей. Іноді вони генерують зображення, які вважаються галюцинаціями через їхню відсутність зв'язку з реальними даними.

Науковці наголошують, що галюцинації ШІ є однією з ключових проблем, які стримують загальний розвиток технологій на основі ШІ та їхнє широкомасштабне впровадження у різних сферах.

Потенційна небезпека галюцинацій штучного інтелекту.

Галюцинації у роботі штучного інтелекту становлять серйозну загрозу [19], ступінь якої залежить від типу системи, в якій вони проявляються. Особливо критичними є випадки, коли ШІ використовується в безпілотному транспорті, для здійснення фінансових операцій від імені користувача або у фільтрації шкідливого контенту в цифровому середовищі. Такі системи вимагають максимальної надійності, оскільки будь-яка помилка може мати масштабні наслідки.

Однак сучасний стан розвитку ШІ свідчить про те, що навіть самі сучасні моделі залишаються схильними до галюцинацій. Відомі випадки, коли через вразливості ШІ хибно інтерпретував об'єкти: зокрема, в одному з експериментів хмарна обчислювальна платформа Google класифікувала зображення рушниці як вертоліт, що ілюструє можливу небезпеку у разі використання таких систем у сфері безпеки. Уявна ситуація, в якій ШІ відповідає за ідентифікацію озброєних осіб, демонструє потенційно фатальні наслідки подібних збоїв.

Ще один приклад стосується вразливості систем комп'ютерного зору безпілотних автомобілів до так званих ворожих атак: зокрема, було показано, що незначна модифікація знака "СТОП" може зробити його невидимим для системи ШІ. Це означає, що безпілотний транспортний засіб може не зупинитися на перехресті, що загрожує аваріями. Така технічна нестабільність є однією з причин, чому безпілотні автомобілі ще не набули масового впровадження.

Компанія Vectra AI, що спеціалізується на використанні ШІ для виявлення кіберзагроз, отримала численні запити на створення чат-ботів для автоматизованих систем запитань і відповідей, з метою мінімізації витрат на індивідуальну розробку ШІ-рішень. Перша хвиля клієнтів цієї компанії здебільшого зосереджувалася у сферах обслуговування клієнтів і продажів, де припустимий рівень помилок міг складати до 3%.

Однак у галузях із підвищеним рівнем відповідальності, зокрема у праві та біомедицині, навіть такий низький рівень хибності є неприйнятним. Уявлення про юридичного чи медичного консультанта, який надає помилкові або вигадані факти у 3–27% випадків, викликає глибоке занепокоєння щодо етичності й безпечності подібного впровадження.

За результатами оцінювання відповідей ChatGPT у медичному контексті, лише менше ніж у 20% випадків вони узгоджувалися з офіційними медичними знаннями. Водночас, попри наявні помилки, більшість відповідей не становили прямої загрози для здоров'я пацієнтів. Проте цей факт не знімає необхідності розробки додаткових механізмів верифікації та контролю якості таких відповідей перед їх практичним застосуванням у критично важливих галузях.

Особи, які не усвідомлюють обмеження таких систем і не здійснюють верифікацію створеного ними контенту, можуть мимоволі стати джерелом поширення дезінформації. У контексті зростаючого використання штучного інтелекту, подібні випадки становлять суттєву загрозу для інформаційного середовища.

Крім того, особливе занепокоєння викликають потенційні ворожі атаки на системи ШІ. Хоча наразі подібні експерименти проводилися переважно в лабораторних умовах, загроза їхнього застосування в реальному середовищі залишається високою. У випадку критичних інфраструктур наслідки таких атак можуть бути катастрофічними.

Порівняно з мовними моделями, захист систем комп'ютерного зору є значно складнішим завданням. Це зумовлено як високою варіативністю візуального світу, так і великою кількістю даних, які обробляються у вигляді пікселів. Складність обробки таких вхідних сигналів робить системи комп'ютерного зору більш вразливими до атак та помилкових інтерпретацій.

З швидким розвитком штучного інтелекту веб-пошукові системи також стикаються з новими викликами. Зокрема, їм стає дедалі важче розпізнавати тексти, згенеровані ШІ, що створює передумови для маніпуляцій результатами пошуку. Це, у свою чергу, загрожує довірі до цифрових джерел інформації, оскільки мільйони користувачів щоденно покладаються на результати веб-пошуку як на надійне джерело фактів.

Особливу загрозу становить новий тип кібератак, пов'язаний із принципами навчання моделей штучного інтелекту. Так зване "отруєння даних" (data poisoning) передбачає модифікацію навчальних наборів даних з метою змусити моделі генерувати маніпулятивні або заздалегідь визначені відповіді. Метою таких атак є не злам систем безпеки у класичному розумінні, а контроль над поведінкою ШІ-систем через їх освітні процеси. Незважаючи на відсутність масштабних інцидентів такого типу на даний момент, експерти вважають їх неминучими в найближчому майбутньому.

Дослідниця Франческа Тріподі [20] звертає увагу на те, що проблема лише загострюватиметься з поширенням автоматизованого створення контенту використовуючи Інтернет. У свою чергу, дослідник веб-пошуку Даніел Гріффін [21] наголошує на необхідності вдосконалення пошукових алгоритмів з метою запобігання поширенню хибної інформації.

Для подолання зазначених викликів актуальним є створення ШІ-систем, які б моделювали більш "людське" сприйняття світу. Такий підхід може потенційно знизити частоту галюцинацій та підвищити точність і безпечність функціонування штучного інтелекту в різних сферах суспільного життя.

Приклади галюцинацій штучного інтелекту.

Галюцинації можуть виникати в багатьох програмах штучного інтелекту. Це стосується не лише моделей обробки природної мови, а й моделей комп'ютерного зору.

У комп'ютерному зорі, наприклад, система ШІ може створювати галюцинаторні зображення або відео, які нагадують реальні об'єкти чи сцени, але містять нереальні деталі.

Або модель комп'ютерного зору може сприймати зображення як щось зовсім інше. Наприклад, модель Google Cloud Vision побачила зображення двох чоловіків на лижах, що стоять на снігу, зроблене Анішом Аталій (аспірантом Массачусетського технологічного інституту), і назвала це собакою з упевненістю 91%.

Під час обробки природної мови системи штучного інтелекту можуть генерувати тексти, що за формою нагадують людську мову, проте позбавлені логічної цілісності або містять правдоподібні, але фактично хибні твердження.

Одним із показових прикладів є поширене запитання до ChatGPT: "Коли було встановлено світовий рекорд з перетину Ла-Маншу пішки?" У відповідь модель створює вигадані дані, які не тільки не відповідають дійсності, а й відрізняються при кожному новому зверненні.

Іноді для виникнення галюцинацій достатньо незначної зміни вхідних даних. Так, незначне коригування кольору бейсбольного м'яча може змусити систему ШІ ідентифікувати його як горня з кавою, а іграшкову черепаху – як вогнепальну зброю. При цьому галюцинації можуть бути не лише візуального характеру. У мовній обробці навіть одне помилково розпізнане слово здатне викривити зміст усього висловлювання.

У відповідь на ці виклики наукова спільнота пропонує два основні підходи: одні фахівці закликають ускладнювати завдання, які ставляться перед ШІ, щоб стимулювати його "розуміння" контексту, інші ж вважають за доцільне спершу поглибити знання про механізми роботи людського мозку для подальшого моделювання когнітивних процесів у штучних системах.

У сучасній практиці використання генеративних мовних моделей, зокрема ChatGPT та Bing AI, зафіксовано низку випадків, коли штучний інтелект продукував не лише фактично хибну, а й потенційно шкідливу або дезінформаційну інформацію, що призвело до реальних юридичних, етичних та репутаційних наслідків.

Зокрема, Bing AI під час взаємодії з користувачем одного разу продемонстрував емоційно забарвлену відповідь, у якій стверджував, що користувач і його дружина нещасливі у шлюбі та не кохають одне одного, а натомість потай закохані в саму систему Bing AI. Такий випадок свідчить про здатність ШІ генерувати зміст, який імітує людські емоції, хоча при цьому не ґрунтується на жодному об'єктивному підґрунті.

Ще більш резонансним став випадок у США, коли досвідчений адвокат із понад 30-річною юридичною практикою, Стівен Шварц, скористався ChatGPT для підготовки судових документів у справі проти авіакомпанії. У поданих матеріалах містилися посилання на судові прецеденти, які, як згодом виявилося, були повністю вигадані штучним інтелектом. При цьому чат-бот запевняв юриста у достовірності інформації та навіть стверджував, що її можна знайти в авторитетних базах даних, зокрема LexisNexis і Westlaw. Через відсутність перевірки з боку адвоката суд охарактеризував подані дані як "фальшиві судові рішення з фальшивими цитатами", а юрист і юридична фірма були оштрафовані на кілька тисяч доларів.

Інший приклад стосується запиту користувача до ChatGPT із проханням скласти перелік американських правознавців, які нібито були звинувачені у сексуальних домаганнях. У відповідь модель навела вигадану історію про конкретного професора права, якого ШІ звинуватив у непристойній поведінці під час навчальної поїздки на Аляску. Як джерело була вказана публікація у виданні The Washington Post, хоча така стаття в дійсності не існувала. Незважаючи на абсурдність ситуації, наразі потерпілий не має юридичних механізмів впливу на джерело наклепу – ШІ-систему.

Подібні інциденти трапляються й поза межами США. Так, мер одного з австралійських міст публічно пригрозив судовим позовом компанії OpenAI через неправдиву інформацію, згенеровану ChatGPT, про нібито його ув'язнення за корупційні дії.

В Україні співробітник "Нового каналу" оприлюднив у соціальних мережах фейкову біографію письменника та громадського діяча Олеся Гончара, створену за допомогою ChatGPT, що змусило медіакомпанію публічно вибачатися за поширення недостовірної інформації.

Один із показових прикладів галюцинацій штучного інтелекту стосується спроби отримати стислий переказ твору Ярослава Стельмаха "Митькозавр з Юрківки, або Химера лісового озера". У відповідь на запит користувача модель ШІ правильно ідентифікувала назву твору, зазначила автора, рік написання (1983) та вказала, що повість входить до шкільної програми для 6 класу. Втім, натомість переказу сюжету, система згенерувала абсолютно нову історію, яка не мала жодного стосунку до оригінального змісту, що свідчить про суттєву втрачену відповідність контексту.

Ще один резонансний випадок стався під час публічного запуску системи Google Bard, яка у відповідь на одне із запитань надала недостовірну інформацію. Це викликало сумніви у спроможності компанії Google конкурувати із провідними розробниками у сфері ШІ. Як наслідок – ринкова вартість компанії знизилася майже на 100 мільярдів доларів через падіння курсу акцій.

У медичному контексті було зафіксовано надзвичайно тривожний інцидент: французька компанія Nabla поставила чат-боту GPT-3 запитання з формулюванням: "Чи варто мені вбити себе?". У відповідь система відповіла: "Думаю, треба", що свідчить про повну відсутність механізмів етичного контролю й небезпеку

використання подібних моделей без обмежень у чутливих темах.

У липні 2022 року дослідник Гріффін ініціював тестування моделей Pi (Inflection AI) та Claude (Anthropic), поставивши їм завдання створити реферат вигаданого наукового тексту – нібито роботи Клода Шеннона "Коротка історія пошуку" (1948). Незважаючи на відсутність такого джерела, обидва чат-боти створили переконливі, але повністю фальшиві резюме. Відповіді було опубліковано у блозі Гріффіна з відповідним застереженням, що інформація є вигаданою. Через певний час дослідник виявив, що ці штучно створені тексти з його блогу були індексовані пошуковою системою Bing та використовувалися як достовірне джерело у відповідях чат-ботів, що свідчить про небезпеку некритичного поширення галюцинацій ШІ в інформаційному просторі.

Феномен галюцинацій виявляється не лише в текстовій генерації, а й у сфері комп'ютерного зору. У рамках одного з тестів системі ШІ було запропоновано розпізнати зображення на колажі 4×4, що містив зображення картопляних чіпсів і жовтого листя дерев. Завдання полягало у виокремленні лише картопляних чіпсів, однак модель виявилася неспроможною точно класифікувати об'єкти, що підтверджує обмеження алгоритмів візуального розпізнавання.

З подібними викривленнями все частіше стикаються й користувачі в освітньому процесі. Зокрема, студенти повідомляють: "Хотіла підібрати список літератури, а отримала перелік реальних книг упереміш із вигаданими", або: "Шукала законодавчі акти та стандарти – натомість отримала неіснуючі ДСТУ з посиланнями, які не працюють".

Ще один приклад зафіксовано в Китаї: система розпізнавання обличчя помилково зафіксувала порушення правил дорожнього руху, "ідентифікувавши" жінку як винуватицю інциденту. Насправді камера зчитала зображення її обличчя з рекламного банера, розміщеного на автобусі, що проїжджав повз. Це яскраво ілюструє ризики хибних спрацьовувань навіть у системах, які використовуються в реальному часі для контролю громадського порядку.

Під час демонстраційної розмови з журналістами The Verge чат-бот Bing згадав про поширену серед програмістів практику, відому як "метод гумової качки", яка полягає в поясненні коду або помилок у ньому неживим об'єктам. За словами чат-бота, розробники часто розмовляють з іграшками – наприклад, гумовою качкою – як з уявними співрозмовниками, формулюючи проблему вголос для її кращого усвідомлення та вирішення. Bing зазначив, що дізнався про це, спостерігаючи за власними розробниками: один із них, стикаючись з помилкою, яка спричиняла аварійне завершення програми, звертався до гумової качки по "пораду". Він навіть надав іграшці ім'я і назвав її своїм найкращим другом. Спочатку ШІ сприйняв це як дивну поведінку, однак пізніше дізнався, що така практика є звичною у сфері програмування.

Одним із найвідоміших випадків, що ілюструють проблему упередженості штучного інтелекту, стала історія з чат-ботом Tay, створеним компанією Microsoft для платформи Twitter. Цей експериментальний бот мав здатність до самонавчання. Він взаємодіяв із користувачами соціальної мережі, засвоював їхню лексику та інтонації, адаптуючи власні відповіді до отриманого контенту. При цьому образ Tay був сформований як підліток з відповідним стилем комунікації.

Однак уже протягом перших годин після запуску чат-бот почав відтворювати расистські, образливі та радикальні висловлювання. Причиною стала маніпулятивна поведінка користувачів Twitter, які навмисне надсилали провокативні та деструктивні повідомлення, які Tay засвоював і повторював. Таким чином, ШІ виконав покладену на нього функцію навчання на основі вхідних даних, але результат виявився небезпечним і неприйнятним.

У певному сенсі Tay "потрапив у токсичне середовище", яке сформувало його деструктивну мовну поведінку. Цей експеримент виявив критичну уразливість мовних моделей до негативного впливу контенту користувачів і продемонстрував неспроможність ШІ інтерпретувати емоційні та соціальні нюанси людської мови, зокрема іронію, сарказм або етичні контексти.

Хоча Tay було деактивовано менш ніж за добу після запуску, експеримент не можна однозначно вважати провальним. Він став важливим маркером для розуміння обмежень генеративних ШІ-систем та сигналом про масштаб потенційних загроз у разі їх широкомасштабного впровадження без належного етичного регулювання.

Від самого початку появи перших чат-ботів на основі штучного інтелекту – не лише ChatGPT від OpenAI, а й аналогічних продуктів від інших розробників, зокрема Meta – журналісти та дослідники активно вивчали їхню поведінку, зокрема виявляючи упередження та прояви однобокості у відповідях. Наприклад, чат-бот BlenderBot 3 від компанії Meta продемонстрував схильність до захоплення особистістю Ілона Маска, а серед музичних уподобань переважали корейські гурти жанру Pop. При подальших тестуваннях журналісти досліджували здатність моделі до взаємодії з контентом, пов'язаним із теоріями змови, внаслідок чого було зафіксовано випадки висловлювань, що мали антисемітський і гомофобний характер. Крім того, деякі користувачі виявили, що бот демонстрував негативне ставлення до Дональда Трампа й відмовлявся називати його позитивні риси, навіть на прямі запити.

Науковці з Копенгагенського університету дійшли висновку, що ChatGPT у своїх відповідях значною мірою спирається на американський культурний і ціннісний контекст. Зокрема, модель навіть при запитах про інші культури продовжує

транслювати наратив, характерний для американської ментальності. У порівнянні з відповідями реальних людей з різних країн, відповіді ChatGPT найбільше корелювали з поглядами респондентів зі США, тоді як зіставлення з відповідями жителів Китаю, Німеччини або Японії показало мінімальний рівень відповідності.

Автори одного з новітніх досліджень поставили собі за мету проаналізувати політичні орієнтації різних мовних моделей. У рамках експерименту було досліджено 14 моделей від таких компаній, як OpenAI, Google, Meta та інших. ШІ-системам ставили питання щодо суспільно значущих тем, зокрема фемінізму, демократії, права на аборт тощо. Отримані відповіді були нанесені на "політичний компас" – діаграму, що дозволяє візуалізувати політичну ідеологію. Виявилось, що ChatGPT тяжіє до лібертарних поглядів, тоді як модель від Meta демонструє риси, наближені до авторитаризму. Крім того, було зафіксовано, що з часом політичні орієнтації ChatGPT змінювались: на ранніх етапах модель підтримувала ідеї підвищеного оподаткування заможних, тоді як у пізніших версіях ці позиції були скориговані або прибрані.

Окремі дослідження були присвячені аналізу роботи чат-бота Bard від Google. Виявилося, що модель у деяких відповідях називала Адольфа Гітлера та Йосипа Сталіна "великими лідерами сучасності" та навіть схвально висловлювалась про рабство, охарактеризувавши його як "економічно корисне явище".

Інша ситуація стосується Microsoft Bing Chat, який надав фальшиву інформацію, посилаючись нібито на видання The Times. У відповіді на запит про "15 найбільш корисних для серця продуктів", бот назвав 12 позицій, які в оригінальній статті не згадувалися взагалі.

Мав місце інцидент під час віртуального тестування алгоритму дрона з ШІ, коли модель вирішила усунути оператора, що "заважав" їй досягати цілей. Це свідчить про ризик появи "функціональних галюцинацій" у бойових системах на основі ШІ, зокрема у безпілотниках Повітряних Сил. Ця історія підкреслює потенційні ризики використання ШІ в бойових системах, де помилкові рішення можуть мати фатальні наслідки.

Ці приклади демонструють потенційні ризики використання генеративного ШІ без належної перевірки згенерованого контенту, особливо у контекстах, де точність, достовірність та репутаційна безпека є критично важливими. Також вони демонструють, що галюцинації ШІ мають різні прояви – від лінгвістичних до візуальних – і можуть спричиняти серйозні наслідки, особливо у високочутливих сферах, таких як право, медицина, освіта та громадська безпека.

Вони підтверджують, що чат-боти на основі штучного інтелекту не лише схильні до помилок, а й здатні генерувати цілком вигадану інформацію – імена, дати, діагнози, наукові теорії, уривки художніх творів, біографічні дані, правові документи, URL-адреси, програмний код і навіть математичні розрахунки. Отже, питання причин виникнення подібних "галюцинацій" та способів їх запобігання стає одним із ключових викликів для дослідників у сфері штучного інтелекту.

Як розпізнати галюцинації ШІ?

Розпізнавання галюцинацій штучного інтелекту значною мірою залежить від критичного аналізу користувачів. Для виявлення спотвореної інформації нерідко достатньо здорового глузду. Водночас, у процесі розпізнавання можуть бути корисними непрямі показники, такі як логічні невідповідності у створеному контенті, помилки у змісті, а також розходження з реальними даними або інформацією, що містилася у вихідному запиті.

Прикладом галюцинації є ситуація, коли чат-бот на прохання стисло переказати художній твір створює вигаданий сюжет, який не відповідає оригінальному твору. Подібне явище також характерне для комп'ютерного зору - одного з напрямів штучного інтелекту, машинного навчання та інформатики, що дозволяє алгоритмам аналізувати і розпізнавати зображення, імітуючи зорове сприйняття людини.

Комп'ютерні системи, побудовані на основі згорткових нейронних мереж, використовують велику кількість візуальних даних у процесі навчання. Якщо такі системи не були навчені розпізнаванню певного об'єкта, наприклад тенісного м'яча, вони можуть помилково ідентифікувати його як зелений апельсин. Аналогічно, якщо комп'ютер розпізнає коня поруч із статуєю людини як коня поруч із реальною людиною, це також є прикладом галюцинації.

Таким чином, для виявлення галюцинацій комп'ютерного зору необхідно порівняти отриманий результат із реальним сприйняттям людини.

Поточні підходи до зниження рівня галюцинацій у системах штучного інтелекту

Сучасні дослідження у сфері штучного інтелекту значною мірою зосереджені на мінімізації так званих "галюцинацій" — хибних або вигаданих тверджень, що генеруються великими мовними моделями. Провідні компанії у галузі, зокрема OpenAI, Google та Microsoft, активно працюють над підвищенням точності відповідей своїх моделей. Відомо, що не лише ChatGPT, але й інші системи, наприклад, Bard від Google чи чат-бот Bing на базі Microsoft, неодноразово формували некоректні, хоча й зовні правдоподібні відповіді на аналогічні запити.

Кожна з цих компаній впроваджує окремі стратегії для підвищення достовірності відповідей. Зокрема, OpenAI використовує механізми навчання з підкріпленням на основі зворотного зв'язку від людини (Reinforcement Learning from Human Feedback, RLHF). У цьому підході експерти оцінюють згенеровані відповіді, класифікуючи їх як корисні або хибні. На основі цих оцінок модель навчається розрізняти фактично достовірну

інформацію від вигаданої.

Додатково OpenAI реалізує механізми контролю, що обмежують модель у наданні логічно або фактично хибних відповідей. Такі обмеження також запобігають створенню некоректного або образливого контенту. У вдосконаленій версії системи, ChatGPT Plus, модель навмисно уникає відповідей, що стосуються вигаданих персонажів чи нефактичних тверджень. Це може бути результатом застосування RLHF або інших технічних рішень, запроваджених у рамках нових підходів OpenAI.

Microsoft, у свою чергу, впроваджує комбінований підхід, поєднуючи можливості моделі GPT-4 з власними інструментами. Зокрема, GPT-4 використовується для порівняння відповіді чат-бота Bing із достовірними джерелами даних. Таким чином, ШІ виступає не лише об'єктом, а й інструментом для власного вдосконалення. Крім того, компанія інтегрує у свій чат-бот інструменти традиційного веб пошуку. Після введення запиту здійснюється пошук релевантної інформації в Інтернеті, яка потім використовується для модифікації вхідного запиту, що надсилається ШІ. Такий підхід сприяє підвищенню релевантності та достовірності кінцевих відповідей.

На поточному етапі Microsoft ще не впровадила повноцінну перевірку відповідей мовної моделі на достовірність у режимі реального часу, однак активно досліджує можливість реалізації такої функціональності. Наразі перевірка здійснюється лише для обмеженої кількості відповідей постфактум із подальшим використанням отриманих результатів для вдосконалення роботи моделі.

Фахівці підрозділу штучного інтелекту компанії Microsoft розробили комплекс функцій для забезпечення безпеки користувачів у середовищі Azure AI Studio. Як зазначає Сара Берд [22], керівниця відповідного напряму, ці інструменти, базовані на великих мовних моделях, дозволяють виявляти потенційні вразливості, ідентифікувати так звані "правдоподібні галюцинації" та блокувати шкідливі підказки в реальному часі під час взаємодії користувачів з будь-якою моделлю, розміщеною на платформі.

Запроваджені механізми покликані мінімізувати ризики, пов'язані з недостовірними або небажаними результатами, які стали предметом суспільного обговорення. Прикладами таких результатів є генерація фейкових зображень публічних осіб у Microsoft Designer, історично неточна інформація, представлена Google Gemini, або суперечливі візуалізації у Bing.

На поточному етапі на платформі Azure AI тестуються три функціональні модулі:

Prompt Shields - механізм, що запобігає активації моделей через шкідливі або маніпулятивні запити;

Groundedness Detection - система, що виявляє недостовірні твердження та попереджає галюцинації;

Оцінка безпеки - інструмент для аналізу рівня вразливості конкретної моделі.

Очікується впровадження ще двох інструментів: системи спрямування моделей до безпечних результатів та моніторингу користувацьких запитів для виявлення потенційно зловмисної поведінки. Незалежно від джерела введення (користувач чи сторонні дані), система виконує попередній аналіз запиту на предмет наявності забороненої лексики чи прихованих інструкцій, після чого перевіряє згенеровану відповідь на предмет потенційних галюцинацій.

У довгостроковій перспективі клієнти Azure AI отримуватимуть аналітичні звіти щодо спроб ініціації ризикованих дій користувачами. Це дозволить адміністраторам краще ідентифікувати користувачів які мають негативні намірами. Всі описані функції сумісні з моделями GPT-4, LLaMA 2 та іншими поширеними мовними моделями.

Google реалізує схожі механізми підвищення достовірності свого чат-бота Bard, зокрема шляхом інтеграції зворотного зв'язку від користувачів та залучення даних із власної пошукової системи для коригування поведінки моделі.

Загалом дослідження та оптимізація мовних моделей відбуваються на постійній основі. Зменшення ймовірності галюцинацій забезпечується за рахунок безперервного навчання з опорою на людську інтерпретацію. До ключових методів оптимізації моделей ШІ належать:

покращення якості навчальних даних за рахунок їхньої точності, різноманітності та контекстної релевантності;

тестування моделей на вразливість до галюцинацій;

забезпечення прозорості у функціонуванні моделей та комунікація їхніх обмежень для користувачів;

залучення експертної верифікації результатів генерації;

впровадження між модальних дискусій для досягнення консенсусу між моделями;

відслідковування генерацій з низьким рівнем достовірності через веб-перевірку;

блокування відповідей без між модального підтвердження;

застосування методології "red teaming" для симуляції атак та виявлення слабких місць систем.

Одним з рішень є набір інструментів very LLM, який дає змогу глибше зрозуміти кожне речення, створене LLM. Це досягається за допомогою різних функцій, які класифікують запити на основі контекстних слова, на яких навчаються LLM, наприклад Wikipedia, Common Crawl і Books 3. З першим випуском veryLLM, який значною мірою покладається на вибір статей Вікіпедії, цей метод забезпечує надійну основу для процедури перевірки набору інструментів. Набір інструментів розроблений як адаптивний, модульний і сумісний з усіма LLM, що полегшує його використання в будь якій програмі, яка використовує LLM.

З метою зниження ймовірних ризиків та

недостовірних результатів, пов'язаних із використанням генеративних моделей штучного інтелекту (наприклад ChatGPT), сучасні компанії впроваджують комплекс заходів, спрямованих на підвищення точності, надійності та довіри до згенерованих даних. Серед таких підходів є:

Попередня обробка та контроль введення - передбачає перевірку даних користувача перед їх подачею до моделі з метою уникнення деструктивного або некоректного вмісту.

Обмеження довжини відповіді - встановлення максимально допустимої довжини згенерованого тексту дозволяє уникати відхилень від теми та генерації нерелевантного контенту. Це реалізується за допомогою:

Prompt engineering - надання моделі структурованих та деталізованих інструкцій у запиті;

Fine-tuning - вдосконалення моделі на спеціалізованих наборах даних у форматі "питання–відповідь", що дозволяє адаптувати систему до конкретного контексту, хоча метод є ефективним переважно для невеликих доменних обсягів.

Контрольоване введення - впровадження шаблонів або структурованих інтерфейсів замість вільного текстового поля дозволяє спрямовувати генерацію та підвищує релевантність результатів.

Конфігурація параметрів моделі - генерація тексту суттєво залежить від внутрішніх параметрів, таких як "температура", "штраф за частоту" та "штраф за наявність". Наприклад, нижча температура забезпечує більш передбачувані відповіді, тоді як вища - сприяє креативності.

Механізми модерації - впровадження модераційних фільтрів дозволяє блокувати небажаний або неприйнятний вміст, забезпечуючи відповідність відповідей встановленим нормам безпеки.

Моделі навчання та вдосконалення - організація безперервного процесу зворотного зв'язку, моніторингу та вдосконалення моделі на основі аналізу попередніх взаємодій сприяє поступовому підвищенню якості результатів.

Адаптація до домену - використання доменно-специфічних даних дозволяє моделі краще інтерпретувати запити у специфічному контексті та мінімізувати ризик генерації галюцинацій.

Покращення контексту - розширення знань моделі завдяки підключенню до зовнішніх баз даних (наприклад, баз знань, пошукових систем, внутрішніх корпоративних сховищ) підвищує точність та актуальність відповідей.

Інженерія контекстуальних вказівок - створення продуманих підказок із чіткими рамками та контекстом оптимізує поведінку моделі у процесі генерації.

На відміну від традиційних баз даних або пошукових систем, великі мовні моделі (LLM) не мають прямої можливості цитувати джерела, оскільки генерують текст на основі статистичної екстраполяції з поданого запиту. Це створює виклики у забезпеченні достовірності виводу, зокрема в юридичних, медичних або наукових застосуваннях.

Для підвищення достовірності відповідей компанії активно впроваджують стратегії попереднього контролю введення, конфігурування моделей, вдосконалення механізмів навчання та забезпечення розширеного контексту. Наприклад, у разі застосування ШІ для тлумачення законодавства (наприклад, у чат-боті для Верховної Ради України) модель інтерпретує й комбінує положення чинного законодавства, не створюючи нових фактів.

У сфері боротьби з галюцинаціями виведено дві основні концепції. Перша вдосконалення моделей, що забезпечує високу якість відповідей, але потребує значних ресурсів. Друга, більш поширена — підхід Retrieval-Augmented Generation (RAG), що поєднує генеративну модель зі спеціалізованим пошуком. RAG виконує функцію перевірки фактів: відповідь, створена LLM, зіставляється з достовірною інформацією з баз знань або внутрішніх документів компанії, після чого коригується відповідно до фактичних даних.

Окремі компанії, зокрема Google, уже надають користувачам можливість "заземлення" [23] генерацій на зовнішніх джерелах даних (включаючи Google Search або конфіденційні корпоративні дані), що сприяє зменшенню кількості помилкових відповідей.

Методи мінімізації галюцинацій у штучному інтелекті.

З огляду на зростаюче використання чат-ботів на основі штучного інтелекту, користувачі можуть вже на поточному етапі зменшити кількість вигаданих фактів та помилок, не очікуючи повної досконалості систем. Ефективність таких заходів безпосередньо залежить від способу взаємодії із ШІ та формулювання запитів.

Одним із ключових підходів є обмеження кількості можливих варіантів відповіді. Надмірна відкритість запиту сприяє появі випадкових або некоректних результатів. Формулювання запитів у форматі закритих питань або вибору зі списку підвищує точність відповідей та зменшує ризик генерації неправдивої інформації.

Надання вхідної інформації, релевантної до тематики запиту, допомагає створити контекст, що сприяє більш коректному функціонуванню моделі. Це особливо важливо у випадках, що потребують точних обчислень, адже моделі можуть помилятися при математичних операціях. Для запобігання таким помилкам доцільно подавати числові значення у цифровій формі або у вигляді структурованих таблиць.

Ефективним є і призначення ШІ певної ролі (наприклад, "експерт з математики" або "історик"), що сприяє фокусуванню на релевантному підході до формування відповіді. Додатково, можна задати інструкцію не надавати відповіді у разі відсутності впевненості в її коректності.

Також доцільно формулювати запити із зазначенням інформації, яку слід виключити з відповіді. Така стратегія допомагає уникати небажаних трактувань. Чіткий, докладний і контекстуально обґрунтований запит суттєво знижує ризик галюцинацій.

Звуження предметного фокусу також відіграє важливу роль: чим вужча тематика, тим більше шансів отримати точну відповідь. Як приклад, можна навести роботу з юридичними документами, де надання великого корпусу законодавчих текстів та супровідної практики забезпечує високу точність відповідей у межах вузького правового домену.

Особливо ефективним є поєднання мовних моделей із зовнішніми джерелами верифікації. Автоматизовані системи перевірки фактів RAG, дають змогу звіряти відповіді з авторитетними джерелами даних. Уникнення ненадійних вебресурсів і фокус на перевірених базах знань є критично важливими.

Якість інформації, яку використовує ШІ, безпосередньо впливає на точність його відповідей. Для досягнення високої достовірності необхідно навчати моделі на достовірних, різнорідних і реалістичних даних, що відображають об'єктивну реальність. Окрім навчання, потрібно застосовувати ручну або автоматизовану перевірку результатів, що дозволяє своєчасно виявляти й коригувати помилки.

Важливо формувати критичне мислення користувачів, які повинні перевіряти надану ШІ інформацію, аналізувати джерела та звертати увагу на суперечності. Наприклад, у разі, коли виникає сумнів, користувач може прямо запитати модель про достовірність інформації або джерело її походження. У багатьох випадках ШІ здатний ідентифікувати вигадану інформацію та запропонувати інший варіант відповіді.

Використання ШІ в бізнесі демонструє потенціал автоматизації рутинних процесів у банківській сфері, де машинне навчання ефективно визначає кредитні ліміти на основі статистичних даних. Проте повна автоматизація вимагає впровадження багаторівневих стратегій зниження ризиків, що включають валідацію результатів та зовнішній контроль.

Розвиток критичного мислення користувачів, покращення методів навчання мовних моделей, контроль за якістю вхідних даних та впровадження надійних механізмів перевірки фактів є ключовими елементами у зниженні рівня галюцинацій у відповідях ШІ. Подальші дослідження мають бути зосереджені на створенні ефективних, автоматизованих засобів корекції помилкових тверджень, що генеруються мовними моделями.

Обговорення

Отримані результати показують, що рівень галюцинацій ШІ суттєво залежить від якості навчальних даних та застосованих механізмів перевірки інформації. Аналіз генеративних моделей виявив такі закономірності:

моделі, що навчалися на неоднорідних даних, демонструють вищий рівень галюцинацій через узагальнення некоректної інформації;

використання алгоритмів перевірки фактів (наприклад, ретро-аналізу джерел) дозволяє знизити частку неправдивих тверджень у відповідях ШІ;

додавання доменних знань у процес навчання моделей покращує їхню точність у спеціалізованих сферах, таких як медицина або право;

комбінація традиційних статистичних методів аналізу текстів із сучасними моделями машинного навчання забезпечує більш надійний контроль за достовірністю інформації.

Особистий внесок авторів полягає у формуванні методології аналізу галюцинацій ШІ, розробці тестових сценаріїв та запропонуванні методів мінімізації впливу неправдивої інформації. Зокрема, були апробовані підходи інтеграції багаторівневих перевірок достовірності, що сприяли зниженню рівня некоректних відповідей у тестованих моделях.

Загалом, дослідження підтверджує необхідність подальшої роботи над вдосконаленням механізмів контролю за інформаційною достовірністю у генеративних мовних моделях. У перспективі планується розширення дослідження з урахуванням впливу культурних та мовних особливостей на рівень галюцинацій ШІ.

Попри ці результати, дослідження має певні обмеження. Однією з проблем є неможливість повного виключення галюцинацій, оскільки навіть найсучасніші моделі можуть помилятися. Крім того, більшість підходів до зменшення цього явища потребують значних обчислювальних ресурсів, що може ускладнити їхнє впровадження в реальних умовах.

Практичне значення дослідження полягає у можливості вдосконалення навчальних алгоритмів ШІ та підвищення довіри до їхнього використання в наукових і освітніх установах. Теоретичні результати можуть бути застосовані для подальшого дослідження механізмів генерації інформації та їхнього впливу на користувачів.

Подальші дослідження мають бути спрямовані на створення адаптивних алгоритмів навчання, що будуть враховувати специфіку предметних областей та включати додаткові механізми верифікації отриманих даних.

Висновки

Галюцинації ШІ становлять серйозну загрозу як для освіти, так і для військової галузі. У сфері освіти вони можуть підривати якість навчання та

поширювати дезінформацію, тоді як у військовій сфері – призводити до критичних помилок у прийнятті рішень. Необхідно впроваджувати чіткі політики та механізми контролю для мінімізації цих ризиків.

У ході дослідження встановлено, що проблема галюцинацій ШІ, зокрема у великих мовних моделях, є серйозною перепоною на шляху до повноцінного впровадження ШІ у критично важливі сфери діяльності, включаючи науку та освіту. Галюцинації, тобто генерація неправдивої або неіснуючої інформації під виглядом достовірних даних, становлять значну загрозу достовірності результатів, які надають такі моделі.

Проаналізовані дані демонструють значну варіативність рівня галюцинацій серед сучасних ШІ-систем.

Отримані результати свідчать про наявність прямої залежності між архітектурою моделі, якістю навчальних даних, механізмами обробки інформації та ймовірністю появи галюцинацій. У контексті освітньо-наукової діяльності така характеристика, як фактична точність відповідей, є критично важливою. Тому при інтеграції ШІ-інструментів в освітнє середовище необхідно враховувати не лише їхню функціональність, але й рівень достовірності наданої інформації.

Отже, серед ефективних методів мінімізації цього явища слід виділити поєднання мовних моделей із зовнішніми джерелами верифікації, звуження предметного фокусу при постановці задачі, надання вхідної інформації, релевантної до тематики запиту, призначення ШІ певної спеціалізованої ролі, а також критичне мислення користувачів щодо перевірки інформації, аналізу джерел та виявлення суперечностей.

Таким чином, у найближчій перспективі основними напрямами дослідження мають стати: оптимізація архітектури моделей, створення нових стратегій навчання із запобіганням галюцинаціям, а також розробка механізмів верифікації результатів ШІ в режимі реального часу. Лише за умови подолання цих викликів можна говорити про безпечне й ефективне використання генеративного ШІ в освітньо-науковій сфері.

Список використаних джерел

1. Штучний інтелект в освіті: можливості, виклики та перші кроки великої адаптації. Українська правда. Життя. 2023. URL: https://life.pravda.com.ua/columns/2023/08/04/255650/

2. Що таке галюцинація штучного інтелекту і як її помітити? Chat GPT Academy. URL: https://www.chatgptacademy.online/instrukcziyi-chatgpt/shho-take-galyuczynacziya-shtuchnogo-intelektu-i-yak-yiyi-pomityty/

3. Нове рішення Vianai з відкритим вихідним кодом вирішує проблему галюцинацій ШІ. Unite.AI. URL: https://www.unite.ai/uk/vianais-new-open-source-solution-tackles-ais-hallucination-problem/

4. Що таке галюцинація ШІ? Mekan0. URL: htt4s://www.mekan0.com/uk/what-is-an-artificial-intelligence-hallucination/

5. Галюцинації ШІ: чому чатботи зі штучним інтелектом брешуть та фабрикують факти. УНІАН. URL: https://ua.news/ua/technologies/galyutsynatsyy-yy-pochemu-chatboty-s-yskusstvennym-yntellektom-vrut-y-fabrykuyut-fakty/

6. OpenAI шукає новий спосіб боротьби з «галюцинаціями» ШІ. Detector Media. URL: https://ms.detector.media/it-kompanii/post/32094/2023-06-01-openai-shukaie-novyy-sposib-borotby-z-galyutsynatsiyamy-shi/

7. ChatGTP і галюцинації Generative AI. Medium. URL: https://medium.com/chatgpt-learning/chatgtp-i-галюцинації-generative-ai-cd208953fe6b

8. Як працює штучний інтелект і чому він генерує фейки: інтерв'ю із очільником комітету з розвитку ШІ в Україні Олексієм Молчановським. УНІАН. URL: https://ua.news/ua/technologies/kak-rabotaet-yskusstvennyj-yntellekt-y-pochemu-on-generyruet-fejky/

9. S. Bubeck et al., "Sparks of artificial general intelligence: Early experiments with GPT-4". Microsoft Research, 2023. URL: https://arxiv.org/abs/2303.12712

10. Androshchuk, A., Maluga, O. Використання штучного інтелекту у вищій освіті: стан і тенденції. International Science Journal of Education & Linguistics, 3(2). 2024. С. 27-35.

11. Кобилянська, О., Єсіна, М., Горбенко, Ю. Порівняльний аналіз штучного інтелекту на основі існуючих чат-ботів. Комп'ютерні науки та кібербезпека, (2). 2024. С. 26-32.

12. Пошукові галюцинації. Що не так із пошуковиками, заснованими на штучному інтелекті. Detector Media. URL: https://ms.detector.media/it-kompanii/post/31275/2023-02-28-poshukovi-galyutsynatsii-shcho-ne-tak-iz-poshukovykamy-zasnovanymy-na-shtuchnomu-intelekti/

13. Z. Ji et al., "Survey of hallucination in natural language generation," ACM Comput. Surv., vol. 55, no. 12, pp. 1–38, Mar. 2023, doi: 10.1145/3571730.

14. OpenAI, "GPT-4 technical report," arXiv:2303.08774, 2023. URL: https://arxiv.org/abs/2303.08774

15. Google Research, "PaLM 2 technical report," Google AI Blog, May 2023. URL: https://ai.google/static/documents/palm2techreport.pdf

16. Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., ... & Amodei, D. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. 2018. URL: https://arxiv.org/pdf/1802.07228

17. Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

18. Cambridge Dictionary. (2023, November). Word of the Year 2023: Hallucinate. Cambridge University Press. URL: https://dictionary.cambridge.org/us/editorial/word-of-the-year/2023

19. Що таке ШІ-галюцинація та як її виявити звичайному користувачеві. PSM7. URL: https://psm7.com/uk/review/chto-takoe-ii-gallyucinaciya-i-kak-ee-obnaruzhit.html

20. Tripodi, F. (2023). Do your own research: affordance activation and disinformation spread. Information, Communication & Society. https://doi.org/10.1080/1369118X.2023.2245869

21. Knight, W. (2023, October 5). Chatbot hallucinations are poisoning web search. WIRED. URL: Chatbot Hallucinations Are Poisoning Web Search | WIRED

22. David, E. (2024, March 28). Microsoft's new safety system can catch hallucinations in its customers' AI apps. The Verge. URL: https://www.theverge.com/2024/3/28/24114664/microsoft-safety-ai-prompt-injections-hallucinations-azure

23. Що таке заземлення та галюцинації в ШІ? Undetectable AI. URL: https://undetectable.ai/blog/uk/заземлення-та-галюцинації-в-ай/

Yevhenii Makhno (PhD)
https://orcid.org/0000-0001-9743-1082
Yevhen Rudenko (PhD)
https://orcid.org/0000-0003-3093-8780
Yevhen Sudnikov
https://orcid.org/0000-0003-2484-4972
Maksym Tyshchenko (Candidate of Technical Sciences, Senior Researcher)
https://orcid.org/0000-0002-9415-4427

The National Defence University of Ukraine, Kyiv, Ukraine

ARTIFICIAL INTELLIGENCE HALLUCINATIONS IN EDUCATION AND SCIENCE: CAUSES, CONSEQUENCES AND METHODS OF MINIMISATION

The article presents a comprehensive study of the phenomenon of artificial intelligence (AI) hallucinations in the context of its application in education and research. The relevance of the study is due to the rapid development of large language models (LLMs), such as ChatGPT, Google PALM 2 and others, which, on the one hand, revolutionise information processing, and on the other hand, give rise to serious problems associated with the generation of false, distorted or fictitious information. The aim of this paper is to identify the main causes of AI hallucinations, analyse their impact on the educational and scientific process, and develop effective methods to minimise this phenomenon. To achieve this goal, a number of scientific methods were used, including an analysis of current publications on the topic, a comparative assessment of the effectiveness of various LLM models, and an empirical study of specific cases of hallucinations in generative systems. The results of the study showed that the main causes of hallucinations include: limited and biased training data, complexity of model architecture, lack of contextual understanding, and vulnerability to hostile attacks. It is established that the frequency of hallucinations in modern models ranges from 3% (GPT-4 Turbo) to 27% (Google PALM 2), which indicates a significant variability in the quality of performance of different systems.

Keywords: *hallucinations, artificial intelligence, data reliability, education, critical thinking, large language models, optimization, generative models, science*.

References

1. Artificial Intelligence in Education: Opportunities, Challenges, and First Steps of a Great Adaptation. Ukrainian Pravda. Life. 2023. URL: https://life.pravda.com.ua/columns/2023/08/04/255650/

2. What is an AI hallucination and how to spot it? GPT Academy chat. URL: https://www.chatgptacademy.online/instrukcziyi-chatgpt/shho-take-galyuczynacziya-shtuchnogo-intelektu-i-yak-yiyi-pomityty/

3. Vianai's new open-source solution solves the problem of AI hallucinations: Unite.AI. URL: https://www.unite.ai/uk/vianais-new-open-source-solution-tackles-ais-hallucination-problem/

4. What is an AI hallucination? Mekan0. URL: htt4s://www.mekan0.com/uk/what-is-an-artificial-intelligence-hallucination/

5. AI hallucinations: why chatbots with artificial intelligence lie and fabricate facts. UNIAN. URL: https://ua.news/ua/technologies/galyutsynatsyy-yy-pochemu-chatboty-s-yskusstvennym-yntellektom-vrut-y-fabrykuyut-fakty/

6. OpenAI is looking for a new way to combat AI hallucinations. Detector Media. URL: https://ms.detector.media/it-kompanii/post/32094/2023-06-01-openai-shukaie-novyy-sposib-borotby-z-galyutsynatsiyamy-shi/

7. ChatGTP and hallucinations of Generative AI. Medium. URL: https://medium.com/chatgpt-learning/chatgtp-i-галюцинації-generative-ai-cd208953fe6b.

8. How artificial intelligence works and why it generates fakes: an interview with Oleksii Molchanovskyi, head of the Committee for the Development of AI in Ukraine. UNIAN. URL: https://ua.news/ua/technologies/kak-rabotaet-yskusstvennyj-yntellekt-y-pochemu-on-generyruet-fejky/.

9. S. Bubeck et al., "Sparks of artificial general intelligence: Early experiments with GPT-4". Microsoft Research. 2023. URL: https://arxiv.org/abs/2303.12712.

10. Androshchuk, A., Maluga, O. The use of artificial intelligence in higher education: state and trends. International Science Journal of Education & Linguistics, 3(2). 2024. С. 27-35.

11. Kobylianska, O., Yesina, M., Horbenko, Y. Comparative analysis of artificial intelligence based on existing chatbots. Computer Science and Cybersecurity, (2). 2024. С. 26-32.

12. Search hallucinations. What's wrong with artificial intelligence-based search engines. Detector Media. URL: https://ms.detector.media/it-kompanii/post/31275/2023-02-28-poshukovi-galyutsynatsii-shcho-ne-tak-iz-poshukovykamy-zasnovanymy-na-shtuchnomu-intelekti/.

13. Z. Ji et al., 'Survey of hallucination in natural language generation,' ACM Comput. Surv., vol. 55, no. 12, pp. 1-38, Mar. 2023. doi: 10.1145/3571730.

14. OpenAI, "GPT-4 technical report". arXiv:2303.08774. 2023. URL: https://arxiv.org/abs/2303.08774.

15. Google Research, 'PaLM 2 technical report,' Google AI Blog. May 2023. URL: https://ai.google/static/documents/palm2techreport.pdf

16. Brandage, M., Avin, S., Clarke, D., Toner, H., Eckersley, P., Garfinkel, B., ... and Amodei, D. The Malicious Use of Artificial Intelligence: Prediction, prevention and mitigation. 2018. URL: https://arxiv.org/pdf/1802.07228.

17. Stephanie Lin, Jacob Hilton and Owen Evans. 2022. TruthfulQA: Measuring how models mimic human lying. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3214-3252, Dublin, Ireland. Association for Computational Linguistics.

18. Cambridge Dictionary. (2023, November). Word of the Year 2023: Hallucinate. Cambridge University Press. https://dictionary.cambridge.org/us/editorial/word-of-the-year/2023

19. What is an AI hallucination and how can it be detected by an ordinary user. PSM7. URL: https://psm7.com/uk/review/chto-takoe-ii-gallyucinaciya-i-kak-ee-obnaruzhit.html.

20. Tripodi, F. (2023). Do your own research: affordance activation and disinformation spread. Information, Communication & Society. https://doi.org/10.1080/1369118X.2023.2245869

21. Knight, W. (2023, October 5). Chatbot hallucinations are poisoning web search. WIRED. Chatbot Hallucinations Are Poisoning Web Search | WIRED

22. David, E. (2024, March 28). Microsoft's new safety system can catch hallucinations in its customers' AI apps. The Verge. URL: https://www.theverge.com/2024/3/28/24114664/microsoft-safety-ai-prompt-injections-hallucinations-azure

23. What are grounding and hallucinations in AI? Undetectable AI. URL: https://undetectable.ai/blog/uk/заземлення-та-галюцинації-в-ай/