

ВИГАДАНІ РЕАЛЬНОСТІ ШТУЧНОГО ІНТЕЛЕКТУ: ПРИЧИНИ ГАЛЮЦИНАЦІЙ У ВЕЛИКИХ МОВНИХ МОДЕЛЯХ

Вінницький національний технічний університет

Анотація

У роботі досліджено явище галюцинацій у великих мовних моделях штучного інтелекту, проаналізовано фактори, що сприяють їх виникненню, а також методи виявлення та запобігання. Розглянуто важливість критичного підходу до використання штучного інтелекту для підвищення надійності та точності результатів, що генеруються моделями.

Ключові слова: галюцинація, велика мовна модель, штучний інтелект, навчання на даних, обмеження штучного інтелекту.

Abstract

The article introduces the phenomenon of hallucinations in large language models of artificial intelligence, analyzes the factors contributing to their occurrence, as well as methods of detection and prevention. The importance of a critical approach to the use of artificial intelligence to improve the reliability and accuracy of the results generated by models is considered.

Keywords: hallucination, large language model, artificial intelligence, learning from data, limitations of artificial intelligence.

Вступ

У сучасному світі штучний інтелект (ШІ) здобуває все більшу популярність і застосовується в найрізноманітніших сферах діяльності. Однією з найбільш обговорюваних його гілок є технологія великих мовних моделей (LLM), які здатні здійснювати складну обробку природної мови. Ці системи використовуються для створення чат-ботів, автоматичних помічників, а також для розв'язання низки інших задач, пов'язаних із розпізнаванням мови. Проте, незважаючи на вражаючі досягнення, ШІ також має свої обмеження, одним з яких є так звані "галюцинації" [1].

Галюцинації ШІ — це явище, коли модель генерує неправдиву або недостовірну інформацію, яка не відповідає реальним фактам чи контексту. Це викликає серйозні занепокоєння серед користувачів та розробників, оскільки такі помилки можуть призвести до неочікуваних та навіть небезпечних наслідків. Тому важливо розуміти причини цих галюцинацій і шляхи їх мінімізації.

Результати дослідження

Галюцинації штучного інтелекту — це термін, що описує ситуації, коли великі мовні моделі або інші системи ШІ генерують дані, які не відповідають реальності. Такі "вигадані реальності" можуть мати різні форми: від вигаданих фактів до перекручених та суперечливих висновків. Як правило, ці помилки виникають через обмеження алгоритмів або через неповноту даних, на яких моделі навчаються.

Термін "галюцинація" був запозичений з медицини, де він описує стан, коли людина сприймає відсутні чи спотворені образи чи звуки. У контексті ШІ цей термін використовується метафорично, бо модель "бачить" або генерує інформацію, якої не існує в реальному світі.

Однією з основних причин, чому мовні моделі можуть генерувати помилкові або вигадані результати, є недостатність або погана якість даних, на яких вони навчаються. Мовні моделі обробляють величезні обсяги текстової інформації, здебільшого з Інтернету, де є багато неконтрольованих і навіть недостовірних джерел. У зв'язку з цим, коли модель тренується на таких даних, вона може "запам'ятовувати" не лише коректну, але й неправильну або суперечливу

інформацію. Враховуючи, що Інтернет наповнений матеріалами різного рівня достовірності, зокрема застарілими, неперевіреними або суб'єктивними думками, модель може випадково вчитись з таких джерел, що впливає на якість її відповідей. Для прикладу, якщо модель була навчена на статтях, що містять застарілу, неточну або суперечливу інформацію, вона може відтворювати ці помилки при формулюванні відповідей, навіть не усвідомлюючи, що вони є хибними. Крім того, недостатньо перевірені або обмежені дані, які надаються мовним моделям, можуть спричинити появу фальшивих фактів або навіть повністю вигаданих деталей [2].

Ще однією причиною галюцинацій є переобладнання, яке відбувається, коли модель "запам'ятовує" конкретні приклади з навчальних даних, замість того, щоб узагальнювати отриману інформацію. Це явище виникає, коли модель надто сильно орієнтується на окремі фрагменти даних або показники, що зустрічаються в навчальному наборі, і не здатна розпізнати загальні закономірності або важливі аспекти, які дозволяють ефективно працювати з новими запитами. Наприклад, якщо навчальний набір містить велику кількість статей, присвячених лише одній темі або певному набору фактів, модель може "запам'ятовувати" окремі факти, ідеї чи фрагменти інформації. При цьому ці фрагменти можуть бути обмеженими і не обов'язково відображати універсальні або актуальні для всіх ситуацій знання. У такому випадку модель може генерувати відповіді, які будуть не зовсім правильними або навіть містити неправдиву інформацію, адже вона не здатна правильно застосувати ці факти в контексті іншого запиту або ситуації. Це може призвести до генерації не зовсім правильних або навіть фальшивих відповідей [3].

Моделі ШІ намагаються створити відповіді, ґрунтуючись на доступних їм даних. Однак іноді, коли модель не має достатньо контексту або інформації для створення точної відповіді, вона може згенерувати неповну або вигадану інформацію. Якщо вхідні дані є обмеженими або нечіткими, модель може "заповнити прогалини", використовуючи найближчі можливі варіанти, навіть якщо ці варіанти не є фактичними чи доречними. Цей процес нагадує творчі методи мислення, коли модель намагається "згенерувати" відповідь за допомогою логіки, проте вона може "згущувати" або виправдовувати свої відповіді, створюючи вигадані елементи.

В ході дослідження розглянуто приклад розв'язування логічного завдання, яке на перший погляд може здатися простим, але насправді вимагає уважного аналізу умов задачі та правильної інтерпретації кожного елементу, щоб уникнути помилок у вирішенні. Задача полягає в тому, щоб визначити загальну кількість яблук на трьох березах, враховуючи кількість гілок різного розміру та яблук на кожній з них (рис. 1).

Зросло 3 берези. На кожній березі по 7 великих гілок. На кожній великій гілці по 7 маленьких гілок. На кожній маленькій гілці – по 3 яблука. Скільки всього яблук росло на дереві?

Щоб знайти кількість яблук на деревах, треба виконати кілька етапів обчислень:

1. **Кількість великих гілок на всіх деревах:** На кожній з 3 беріз по 7 великих гілок, тому на всіх деревах:
$$3 \times 7 = 21 \text{ велика гілка}$$
2. **Кількість маленьких гілок на всіх великих гілках:** На кожній великій гілці по 7 маленьких, тому на всіх великих гілках:
$$21 \times 7 = 147 \text{ маленьких гілок}$$
3. **Кількість яблук на всіх маленьких гілках:** На кожній маленькій гілці по 3 яблука, отже, на всіх маленьких гілках буде:
$$147 \times 3 = 441 \text{ яблуко}$$

Отже, на деревах росло всього **441 яблуко**.

Рис. 1. Приклад галюцинацій у великих мовних моделях при вирішенні логічних задач

Галюцинації в ШІ часто виникають через відсутність узгодженості між частинами інформації, яку генерує модель. Якщо модель не може адекватно узгодити різні елементи тексту, то результат може виглядати як абсурд чи плутанина. Наприклад, деякі моделі можуть згенерувати відповідь, яка начебто має сенс, але при більш детальному аналізі виявляється суперечливою або неможливою з реальними

фактами.

ШІ також може створювати вигадані відповіді через надмірне узагальнення. Моделі ШІ часто намагаються створити відповіді на основі спільних закономірностей у даних. Проте іноді ці закономірності є неточними або занадто загальними, що може призвести до створення не зовсім коректної чи неіснуючої інформації. Незважаючи на те, що галюцинації є серйозною проблемою для великих мовних моделей, вони не завжди важко виявляються, і навіть при наявності цієї проблеми можна здійснювати певні кроки для їх виявлення та запобігання (рис. 2).

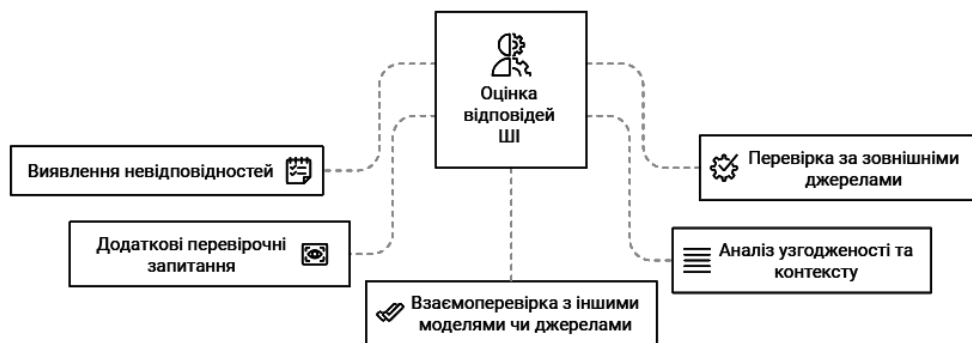


Рис. 2. Основні підходи до виявлення та запобігання галюцинаціям у великих мовних моделях

Оскільки великі мовні моделі часто не мають справжнього семантичного розуміння і працюють на основі статистичних зв'язків між словами та фразами, дуже важливо уважно звертати увагу на смислові невідповідності в їхніх відповідях. Якщо відповідь моделі виглядає непослідовною або нелогічною, це може бути ознакою галюцинації, коли модель неправдиво чи випадково формулює інформацію, яка не відповідає дійсності [4].

Ще одним важливим кроком для виявлення галюцинацій є перевірка наданої моделлю інформації. Якщо відповідь виглядає підозрілою або сумнівною, рекомендується звернутися до надійних, перевірених джерел для підтвердження або спростування отриманої інформації. Це не тільки допоможе уникнути можливих помилок, але й дозволить прийняти обґрунтовані рішення, оскільки неправильна або вигадана інформація може призвести до серйозних наслідків, особливо у важливих сферах, таких як медицина, право чи фінанси.

Галюцинації також можна помітити через порушення узгодженості або контексту. Наприклад, якщо відповідь моделі суперечить поточній розмові, вже сказаному або загальним знанням, це має бути сигналом для перевірки достовірності та уточнення відповіді. У таких випадках важливо звернути увагу на логіку розвитку дискусії і зрозуміти, чи не виникає конфлікту між заявленими фактами.

Для того, щоб оцінити точність наданої інформації, корисно ставити додаткові запитання або перевіряти модель за допомогою різних варіантів формулювань. Якщо модель не може надати послідовної та логічної відповіді, це може свідчити про наявність галюцинації. Також можна застосувати техніку суперечливого тесту, коли навмисно ставляться суперечливі чи парадоксальні запитання для перевірки, як модель реагує на них, що дозволяє виявити її слабкі місця.

Ніколи не варто покладатися лише на одну модель чи джерело інформації, оскільки це може бути небезпечно в разі виникнення галюцинацій або інших неточностей. Взаємоперевірка результатів різними моделями та джерелами допомагає значно зменшити ризик виникнення галюцинацій і підвищує надійність отриманих даних, а також дозволяє користувачам отримати більш комплексну та об'єктивну картину [5-6].

Найважливіший крок для запобігання галюцинаціям — це усвідомлення обмежень великих мовних моделей. Розуміння того, що штучний інтелект не має реального розуміння світу і базується лише на статистичних методах обробки тексту, дозволяє краще зрозуміти, чому моделі можуть створювати неточні або вигадані відповіді. Це усвідомлення дає змогу користувачам підходити до використання таких технологій з більшою обережністю та критичним мисленням, що є запорукою правильного та безпечного застосування ШІ у різних сферах.

Отже, правильне використання технологій ШІ, з урахуванням їхніх обмежень, може значно знизити негативний вплив галюцинацій і забезпечити більш надійний і безпечний процес прийняття рішень.

Висновки

Таким чином, галюцинації в ШІ — це важливий аспект, який потрібно враховувати при використанні великих мовних моделей. Розуміння причин виникнення цих помилок та методів їх виявлення і запобігання дозволяє користувачам використовувати ШІ з більшою обережністю та ефективністю. Незважаючи на свої недоліки, технології штучного інтелекту продовжують розвиватися і вдосконалюватися, що дозволяє з часом зменшувати частоту таких помилок і підвищувати точність моделей.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. IBM. What Are AI Hallucinations? . IBM - United States. URL: <https://www.ibm.com/topics/ai-hallucinations> (дата звернення: 10.11.2024).
2. What are AI hallucinations? | Google Cloud. URL: <https://cloud.google.com/discover/what-are-ai-hallucinations> (дата звернення: 12.11.2024).
3. Maggiolo G. Detecting AI Hallucinations:Identifying False Info from Neural Networks. URL: <https://blog.pigro.ai/en/can-ai-experience-hallucinations-how-to-identify-false-information-generated-by-neural-networks> (дата звернення: 08.11.2024).
4. Що таке галюцинації штучного інтелекту та як компанії вирішують цю проблему. Домени – перевірка та реєстрація доменів в Україні | Імена.ua. URL: <https://www.imena.ua/blog/hallucinations-of-artificial-intelligence/> (дата звернення: 12.11.2024).
5. Що таке галюцинація штучного інтелекту і як її помітити? - Академія ChatGPT Academy - навчання по ChatGPT. URL: <https://www.chatgptacademy.online/instrukcziyi-chatgpt/shho-take-galyuczynacziya-shtuchnogo-intelektu-i-yak-viyi-pomityty/> (дата звернення: 13.11.2024).
6. Крамаренко О. Що таке ШІ-галюцинація та як її виявити звичайному користувачеві. PaySpace Magazine. URL: <https://psm7.com/uk/review/chto-takoe-ii-gallyucinaciya-i-kak-ee-obnaruzhit.html> (дата звернення: 13.11.2024).

Василина Анастасія Василівна – студентка групи 2БС-226, факультет інформаційних технологій та комп’ютерної інженерії, Вінницький національний технічний університет, м. Вінниця, e-mail: nstvsln@gmail.com.

Гарнага Володимир Анатолійович — доцент кафедри захисту інформації, Вінницький національний технічний університет, м. Вінниця, e-mail: garnaga.volodymyr@vntu.edu.ua.

Vasylyna Anastasia Vasylivna - Faculty of Information Technologies and Computer Engineering, Vinnytsia National Technical University, Vinnytsia.

Harnaha Volodymyr Anatoliyovych — Associate Professor of Information Protection, Vinnytsia National Technical University, Vinnytsia.