світової війни, а пізніше поглибились, також про себе дав знати ленд-ліз, післявоєнна позика США, ще більша робила складнішу ситуацію для Британії, оскільки стартова позика за ленд-лізом становила понад 1 мільярд фунтів стерлінгів, полегшення було розтягнуто на 50 платежів які здійснювались, що року починаючи з 1951 року, останній платіж у розмірі 83.3 мільйона доларів у 2006 році. [2,9]

Міцні зв'язки між Великої Британії та Сполучених Штатах Америки лягли в основу заснування ООН і НАТО, якщо брати ООН, то США і Британія з 1945 року, займають головні місця в раді безпеки, щоб забезпечити світовий баланс і не допустити нової світової війни, також НАТО, Велика Британія і США їхня співпраця в часи Другої світової війни лягло в фундамент, функціонуванню НАТО в 1949 році, щоб рятувати країни Європи, від Радянського союзу. США і Велико Британія і інші країни НАТО, закріпили в основі політиці НАТО статтю 5, що означає, що вразі агресії з боку СРСР, кожна країна буде вважати, що напад на неї. Відносини США і Британії, лягли в основу відновленню і безпеки західних демократії від агресії в другій світовій війні, та можливій агресії з боку СРСР. [11,12]

ЛІТЕРАТУРА

1. https://history.state.gov/milestones/1937-1945/atlantic-conf
2. https://www.rand.org/pubs/commentary/2024/07/britain-and-america-why-so-special.html
3. https://www.nationalww2museum.org/war/articles/lend-lease-eastern-front
4. https://history.state.gov/milestones/1937-1945/lend-lease
5. https://encyclopedia.ushmm.org/content/en/article/lend-lease
6. https://www.archives.gov/milestone-documents/lend-lease-act
7. https://www.fdrlibrary.org/lend-lease
8. https://en.wikipedia.org/wiki/United_Kingdom%E2%80%93United_States_relations_in_World_War_II
9. https://cpcml.ca/Tmlm2022/Articles/M520067.HTM
10. https://commonslibrary.parliament.uk/research-briefings/cdp-2025-0027/
11. https://www.archives.gov/milestone-documents/united-nations-charter
12. https://diplomatie.belgium.be/en/policy/policy-areas/peace-and-security/international-organisations/north-atlantic-treaty-organisation-nato

Олексійко Юрій
Науковий керівник – доц. Морська Наталія

ФІЛОСОФІЯ ШІ: ГАЛЮЦИНАЦІЇ, УПЕРЕДЖЕННЯ, РИЗИКИ ТА СТРАТЕГІЇ ЇХ МІНІМІЗАЦІЇ

Сучасний генеративний штучний інтелект – це не просто новий тип «розумних» інструментів – це певний клас моделей, навчених створювати абсолютно новий контент. Вони вчаться та аналізують величезні масиви даних, знаходячи та відтворюючи складні людські патерни. Тому, у їхній роботі є певна особливість – феномен «галюцинації». Модель інколи може видавати дуже правдоподібну, але по свої суті неправдиву інформацію. Беручи до уваги ще те, що моделі іноді можуть втрачати «забувати» інструкції, контекст або плутаються в навчальних даних. Це досить нові явища, які ще не до кінця зрозумілі і досліджені, і саме тому ця тема, є на сьогодні такою актуальною.

Тож чому виникають ці «галюцинації»? Звичайний комп'ютер просто витягує дані з пам'яті. GenAI працює зовсім по іншому. Він не «розуміє» значення нашому розумінні цього слова. Натомість, модель обирає вивчені раніше шаблони і статистично комбінує їх, щоб сформувати нову, унікальну відповідь. Саме цей процес творення, а не копіювання, і призводить до появи результатів, які виглядають логічно, але є хибними у людському розумінні.

Можна провести цікаву паралель з людськими снами. Коли ми спимо, мозок не просто «переглядає» спогади, а перемішує різні фрагменти знань та досвіду, створюючи неможливі сценарії. Вони видаються реальними, хоча переважно є суцільною вигадкою і не спираються на емпіричний досвід. В аналогії GenAI: він не може генерувати коректну відповідь, якщо в його «досвіді» (навчальні дані) просто не існує потрібної інформації. Найглибша проблема полягає в тому, що модель не може усвідомити, коли їй недостатньо знань для коректної відповіді.

Також існує проблема упереджень або «Bias». Ці відповідь зазвичай виглядають занадто абстрактними або дуже з дуже вузьким поглядом. Модель обирає ті варіанти, які вона найчастіше могла зустріти під час свого навчання, не беручи до уваги також менш популярні дані або не такі поширені в її базі, але ці варіанти є також абсолютно логічною альтернативою для відповіді.

Для прикладу, існує «упередження прив'язки» (Anchoring Bias): модель надто сильно зациклюється на початковий запит. Якщо запитати про «найкращого філософа всіх часів», відповідь «Аристотель» може з'явитися просто тому, що перше ж формулювання підштовхує її до «очевидного» вибору, обмежуючи розгляд інших. Інший тип – «упередження через частоту» (Exposure Bias). Тут все набагато простіше: модель обирає найпопулярніше. Якщо ми зробимо запит щоб порадити фільм, то з великою ймовірністю вона може згадати такий фільм як «Титанік», бо цей варіант фільму зустрічався в її навчальних даних дуже часто що і призвело до таких результатів. Зрештою ще існує так звана «систематична помилка вибірки» (Selection Bias). Це може відбуватись, коли такі навчальні дані не є репрезентативними. Для прикладу, модель, яка під час навчання була навчена відповідно тільки на західній літературі, буде ігнорувати погляди з інших культур.

Наслідки такої необ'єктивності, можуть бути дуже серйозними. Приміром, штучні моделі можуть зазвичай асоціювати конкретні професії з відповідною статтю або етнічною групою. Системи для роботи HR, навчені на історичних даних про найм, які можуть бути застарілими в сучасних реаліях, можуть частіше давати керівні посади чоловікам через історичну відповідність. Системи, які вміють розпізнавати обличчя можуть також видавати більше помилок на зображеннях представників етнічних груп, що були погано представлені в навчальних даних моделі. Це призводить не тільки до помилки, а й створює справжню несправедливість.

Що потрібно робити, щоб не потрапити у цю системну пастку. Потрібно зробити певний новий підхід роботи з моделями. Більшість випадків покривають самі творці моделі закладаючи певні обмеження, щоб уникнути банальних етичних чи юридичних проблем. Але з іншого боку,

ми як користувачі ШІ, маємо надавати якомога точніше та конкретні інструкції для роботи. В більшості випадків саме наші неточні інструкції і стають причиною упередженої відповіді.

Але, головна тактика зменшення ризиків утворення «галюцинації» – це підхід «людина в циклі» (Human-in-the-Loop або HITL). Ідея доволі проста: людина має брати безпосередню участь у центральних етапах роботи з ШІ, щоб мати контроль за процесом роботи, перевіряти якість і приймати вирішальні рішення. Мова про повний і безперечний контроль, який би звів нанівець всю ефективність роботи моделі, а про керування у центральних моментах для коригування чи затвердження результатів.

У реальності це має приймати такий вигляд. При створенні контенту AI видає перший варіант відповіді, а вже контролер-людина встановлює тон і перевіряє точність фактів. В наданні послуг клієнту, система видає кілька варіантів, але оператор-людина приймає фінальне рішення, яку саме відповідь надіслати.

Такий підхід дає змогу тримати високу якість, не роблячи компроміс в ефективності роботи. Це дуже важливо там, де помилки або етичні вимоги надто високі і можуть мати високу ціну для, наприклад, медицини, освіти чи при наймі на роботу.

З цього можна зробити певний висновок. Щоб якнайкраще і ефективніше працювати з моделями, нам потрібно зробити свій, новий підхід до надання інструкцій. Потрібно ставитись до систем не як до безпомилкових оракулів, а як до надзвичайно швидких і зручних помічників, які інколи генерують статистично помилкові результати. Коли ми якнайкраще, усвідомлюємо сильні та слабкі сторони моделі, ми можемо як найефективніше використати її .

Головна перевага яку дає AI не в тому, щоб повністю вилучити людину з процесу, а в тому, щоб підсилити її можливості. Для прикладу, викладач може за короткий час створити кілька варіантів тесту за допомогою ШІ, але саме ця людина несе відповідальність за якість і коректність цих створених завдань. Найефективніший результат ми отримаємо лише тоді, коли люди та ШІ будуть доповнювати одна одну, беручи до уваги слабкі сторони у спільній роботі.

ЛІТЕРАТУРА

1. Wikipedia: Generative artificial intelligenceGenerative artificial intelligence. URL: https://en.wikipedia.org/wiki/Generative_artificial_intelligence (дата звернення: 20.10.2025).
2. Medium-стаття про галюцинації в генеративному ШІRibeiro Neto J. A. «ChatGTP і галюцинації Generative AI». URL: https://medium.com/chatgpt-learning/chatgtp-%D1%96-%D0%B3%D0%B0%D0%BB%D1%8E%D1%86%D0%B8%D0%BD%D0%B0%D1%86%D1%96%D1%97-generative-ai-cd208953fe6b
3. (дата звернення: 20.10.2025).
4. Wikipedia: Hallucination (artificial intelligence)Hallucination (artificial intelligence). URL:
5. https://en.wikipedia.org/wiki/Hallucination_(artificial_intelligence)
6. (дата звернення: 20.10.2025).