

Технології штучного інтелекту та безпека даних

Освітня програма «Економіка і аналітика», 1 курс · 4 кредити ECTS

Лекція 4. Надійність генеративного ШІ: чому не можна вірити всьому

Мета лекції: пояснити механізм, через який генеративний ШІ помиляється — не як «баг», а як закономірний наслідок того, як він влаштований — і дати практичні орієнтири для критичного споживання його відповідей.

Зв'язок з попередніми лекціями: лекція 2 показала, що первинна задача ГШІ — генерувати схожий на людський текст, а не істину; лекція 3 показала побічні ефекти масового використання. Сьогодні заглиблюємось у механіку конкретних типів помилок і їх практичні наслідки.

1. Як «думає» ШІ (нагадування)

- ШІ не думає в людському сенсі. У нього немає «бази правильних відповідей» — це не пошукова система і не довідкове бюро.
- Генеративні мовні моделі генерують відповідь, передбачаючи, яке слово статистично найімовірніше буде далі (пор. лекцію 1 — статистичне передбачення наступного слова).
- Ключовий наслідок: модель **завжди** видає якусь відповідь, що звучить правдоподібно, — незалежно від того, чи має вона фактичне підтвердження.

2. Цільова функція генеративних моделей: не правда, а задоволення співрозмовника

- Будь-яка задача оптимізації — це пошук максимально хорошого рішення за заданих обмежень; що саме є «хорошим», визначає **цільова функція**.

Спрощена ієрархія цільової функції генеративної мовної моделі (за пріоритетом):

- 1) Головне — зробити текст максимально подібним на людський.
- 2) Бажано, але не обов'язково — дати правильну відповідь (користувачі іноді перевіряють факти й задоволені більше, коли відповідь правильна).
- 3) Було б непогано — знайти найкращий варіант серед хороших.
- 4) Сказати «не знаю» — останній прийнятний варіант, і лише якщо користувач сам про це прямо просить.

Практичний наслідок: правильність відповіді — це побічний ефект оптимізації, а не первинна мета моделі.

3. Це працює і з людьми

- Психологічний принцип «впевнений тон переконує» діє не лише на сприйняття тексту від ШІ, а й у людському спілкуванні: авторитетна, впевнена подача підвищує довіру незалежно від фактичної точності.
- Це не виправдання для ШІ, а привід усвідомити власну вразливість: ми й одне одного часто оцінюємо за формою (впевненість, стиль), а не за змістом.

- Саме на цю людську схильність «лягає» стиль генерації ГШІ — і тому текст моделі здається переконливим навіть тоді, коли він фактично помилковий.

4. Чому не можна вірити всьому, що каже ШІ

- Прямий наслідок пп. 1–2: модель оптимізована звучати переконливо, а не бути точною.

Головне правило споживання контенту, згенерованого ШІ: впевнений тон ≠ правильна відповідь.

5. Галюцинування

- Якщо ви ставите достатньо вузькоспеціалізоване чи маловідоме запитання, ШІ може «вигадати» правдоподібну відповідь замість того, щоб чесно визнати брак даних — особливо якщо ви прямо не просите модель перевірити факти.
- Механізм: модель «заповнює» відсутню інформацію (назву, дату, джерело, цитату) тим, що статистично найбільш пасує за контекстом — не тому, що свідомо бреше, а тому, що в неї немає вбудованого механізму «я не знаю» (пор. п. 2, пункт 4 ієрархії).

Дидактична примітка: Хороша вправа: попросіть студентів задати ШІ вузькоспеціалізоване запитання зі своєї сфери (курсова, наукова стаття, конкретна цифра) і перевірити відповідь незалежно.

6. Еволюція галюцинування — не все однозначно

- Здавалося б, з розвитком моделей проблема галюцинацій має поступово зникати — але це не завжди так.
- Дослідження (Toraz, Roguin, Gupta et al., *The Lancet*, 2025) проаналізувало 2,5 млн біомедичних наукових статей і зафіксувало зростання частки **сфабрикованих цитувань**: приблизно з 4 на 10 000 статей на початку 2023 року до **57 на 10 000** на початок 2026 року.
- Можливе пояснення: що більше людей використовують ШІ для написання наукових текстів (і що менш ретельно перевіряють згенероване), то більше сфабрикованих посилань потрапляє в публікації — це проблема не тільки технологічна, а й пов'язана зі звичками користувачів.
- Висновок: зростання «формальної» якості моделей (плавність тексту, впевненість тону) не гарантує пропорційного зростання фактичної точності.

7. Упередженість (bias)

- ШІ навчається на даних, які відображають наявні в суспільстві нерівності й стереотипи, тому може відтворювати їх у своїх відповідях чи згенерованих зображеннях.
- Приклад: генерація зображень чи текстів за професійними ролями (наприклад, «секретар», «інженер», «керівник») часто відображає гендерні чи інші стереотипи, наявні в навчальних даних, а не об'єктивну реальність.
- Наслідок для бізнесу: якщо ШІ використовується в кадрових рішеннях, кредитному скорингу чи інших чутливих сферах, таке упередження може призводити до

дискримінаційних результатів — це наскрізна тема, до якої ми повернемось у лекції про право й регулювання ШІ.

8. Дискусійні теми: мова і джерела впливають на відповідь

- Один і той самий запит, поставлений різними мовами або з різним культурним контекстом, може дати відповіді з різним акцентом чи трактуванням — особливо щодо дискусійних історичних або політичних питань.
- Причина: різні мовні сегменти навчальних даних відображають різні національні наративи, джерела й точки зору, домінантні в тому чи іншому інформаційному просторі.
- **Практичний висновок:** для дискусійних тем варто свідомо перевіряти відповідь кількома мовами чи джерелами і не сприймати першу відповідь ШІ як нейтральну «істину в останній інстанції».

9. Ефект чорної скриньки

- Навіть розробники не завжди можуть точно пояснити, чому модель дала саме таку відповідь, — якщо прямо не змусити модель цитувати перевірені джерела.
- Наслідок: не варто сприймати відповідь ШІ як остаточний вердикт, особливо в питаннях здоров'я, права чи фінансів — там, де ціна помилки висока.

10. Дипфейки та дезінформація

- Ті самі технології, що генерують корисний контент, можуть генерувати переконливі фейкові фото, аудіо та відео.
- Ці інструменти можуть потрапити в «погані руки» — для шахрайства, маніпуляції громадською думкою, завдання репутаційної шкоди.
- Це пряма технологічна основа для соціальної інженерії та шахрайства, до яких ми детальніше повернемось у блоці про кібербезпеку.

11. Ключове правило споживання контенту

- Якщо контент (від ШІ чи будь-якого іншого джерела) провокує сильну емоційну реакцію — це сигнал зупинитися й перевірити джерело або знайти незалежне підтвердження, а не одразу довіряти чи ділитися.
- Емоційно «заряджений» контент (обурення, страх, розчулення) — улюблений інструмент і дезінформації, і генеративного ШІ, який оптимізує «залучення читача», а не істину.

Місток до наступної лекції: якщо ми тепер розуміємо, чому і як саме ГШІ помиляється, логічний наступний крок — навчитися формулювати запити так, щоб мінімізувати ці ризики й отримувати від моделі максимально корисний результат. Це тема лекції 5 — промпт-інжиніринг.

Контрольні питання для самоперевірки студентів

- Поясніть своїми словами, чому цільова функція генеративної моделі не збігається з «казати правду».
- Наведіть приклад ситуації, коли ШІ, на вашу думку, найімовірніше «згалюцинує» відповідь.
- Чому кількість сфабрикованих цитувань у наукових статтях зростає, а не спадає з часом, попри покращення моделей?
- Наведіть свій приклад потенційного упередження (bias), яке ШІ міг би відтворити у вашій професійній сфері.
- Чому відповідь ШІ на дискусійне історичне чи політичне питання не варто сприймати як єдину «об'єктивну істину»?
- Сформулюйте «ключове правило споживання контенту» своїми словами і наведіть приклад його застосування.