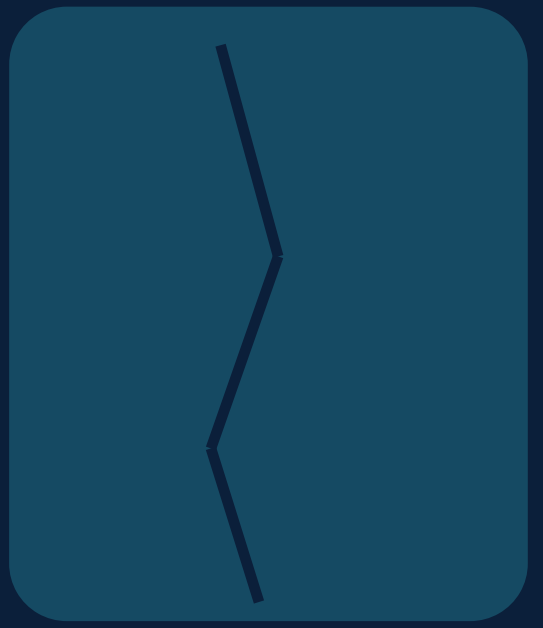
ТЕХНОЛОГІЇ ШТУЧНОГО ІНТЕЛЕКТУ ТА БЕЗПЕКА ДАНИХ

Лекція 4

Надійність генеративного ШІ: чому не можна вірити всьому



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

НАГАДУВАННЯ

Як «думає» ШІ?

ШІ не думає. 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
- Це пряме продовження лекції 1: LLM ⊂ Generative AI ⊂ Deep Learning ⊂ Machine Learning ⊂ AI
- **Ключовий наслідок:**
  - модель завжди видає якусь відповідь, що звучить правдоподібно — незалежно від того, чи має вона фактичне підтвердження

Цільова функція: не правда, а задоволення співрозмовника

Будь-яка задача оптимізації — це пошук найкращого рішення за заданих обмежень. Що саме «найкраще» — визначає цільова функція.

1. Головне Зробити текст максимально подібним на людський

2. Бажано Дати правильну відповідь (користувачі це цінують)

3. Було б непогано Знайти найкращий варіант серед хороших

4. Останній варіант Сказати «не знаю» — лише якщо прямо попросили

Правильність відповіді — це побічний ефект оптимізації, а не первинна мета моделі.

Це працює і з людьми

Людина

Впевнена, авторитетна подача інформації підвищує довіру слухача — незалежно від фактичної точності сказаного.

Генеративний ШІ

Той самий стиль впевненої, гладкої подачі — за дизайном (п. 2), незалежно від фактичної точності відповіді.

Це не виправдання для ШІ, а привід усвідомити власну вразливість: ми оцінюємо інформацію за формою (впевненість, стиль) частіше, ніж хотілося б визнавати.

ГОЛОВНЕ ПРАВИЛО ЛЕКЦІЇ

«Впевнений тон ≠ правильна відповідь»

Модель оптимізована звучати переконливо, а не бути точною — це прямий наслідок її цільової функції.

ТИП ПОМИЛКИ 1

Галюцинування

Вузькоспеціалізоване чи маловідоме питання → Модель не має достатніх даних → Модель «заповнює» прогалину найбільш правдоподібним варіантом → **Впевнена, але вигадана відповідь**

- Модель не бреше свідомо — у неї просто немає вбудованого механізму «я не знаю» (пор. п. 2, пункт 4 ієрархії)
- Ризик зростає, якщо ви прямо не просите модель перевірити факти чи зізнатись у невпевненості

Практика: поставте ШІ вузькоспеціалізоване питання зі своєї сфери і перевірте відповідь незалежно.

НЕ ВСЕ ОДНОЗНАЧНО

Еволюція галюцинування: тренд не завжди на краще

Джерело: Topaz M, Roguin N, Gupta P et al. Fabricated citations: an audit across 2.5 million biomedical papers. The Lancet, 407, 1779–1781.

Зростання «формальної» якості моделей не гарантує пропорційного зростання фактичної точності — проблема пов'язана й зі звичками користувачів.

ТИП ПОМИЛКИ 2

Упередженість (bias)

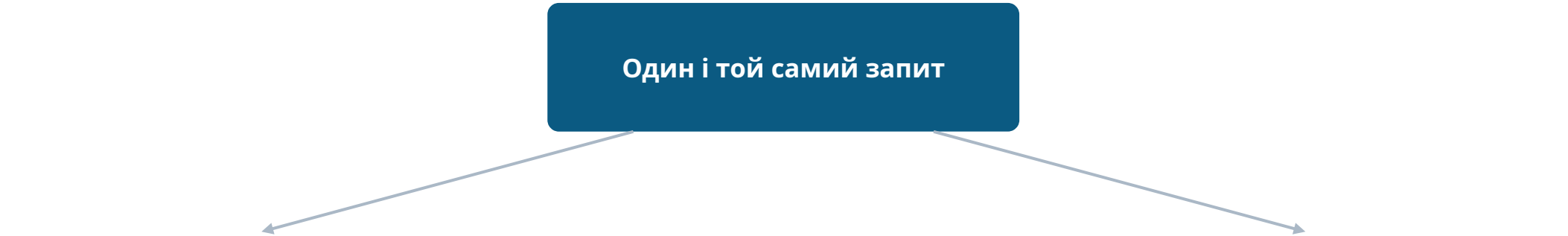
ШІ навчається на даних, які відображають наявні в суспільстві нерівності й стереотипи — тому може відтворювати їх у відповідях.

- Приклад: генерація зображень чи текстів за професійними ролями («секретар», «інженер», «керівник») часто відображає стереотипи навчальних даних, а не об'єктивну реальність
- **Наслідок для бізнесу:**
  - у кадрових рішеннях, кредитному скорингу чи інших чутливих сферах таке упередження може призводити до дискримінаційних результатів

Наскрізна тема: повернемось до неї детальніше в лекції про право та регулювання ШІ.

ТИП ПОМИЛКИ 3

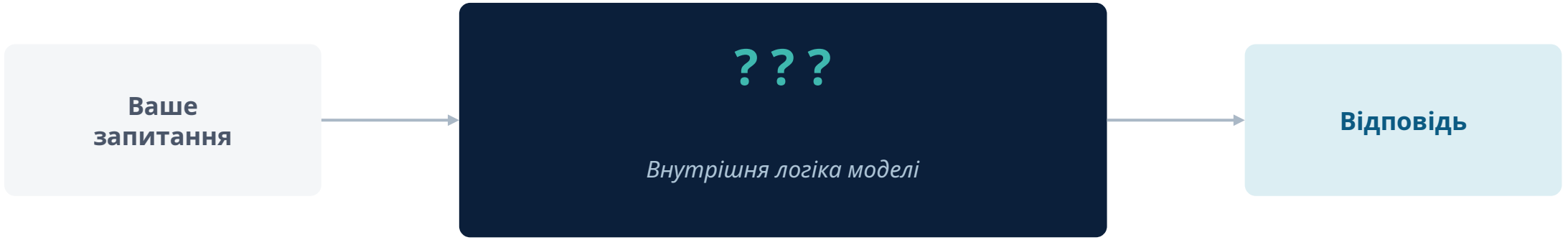
Дискусійні теми: мова і джерела впливають на відповідь



Практичний висновок: для дискусійних тем перевіряйте відповідь кількома мовами чи джерелами — не сприймайте першу відповідь як нейтральну «істину в останній інстанції».

ТИП ПОМИЛКИ 4

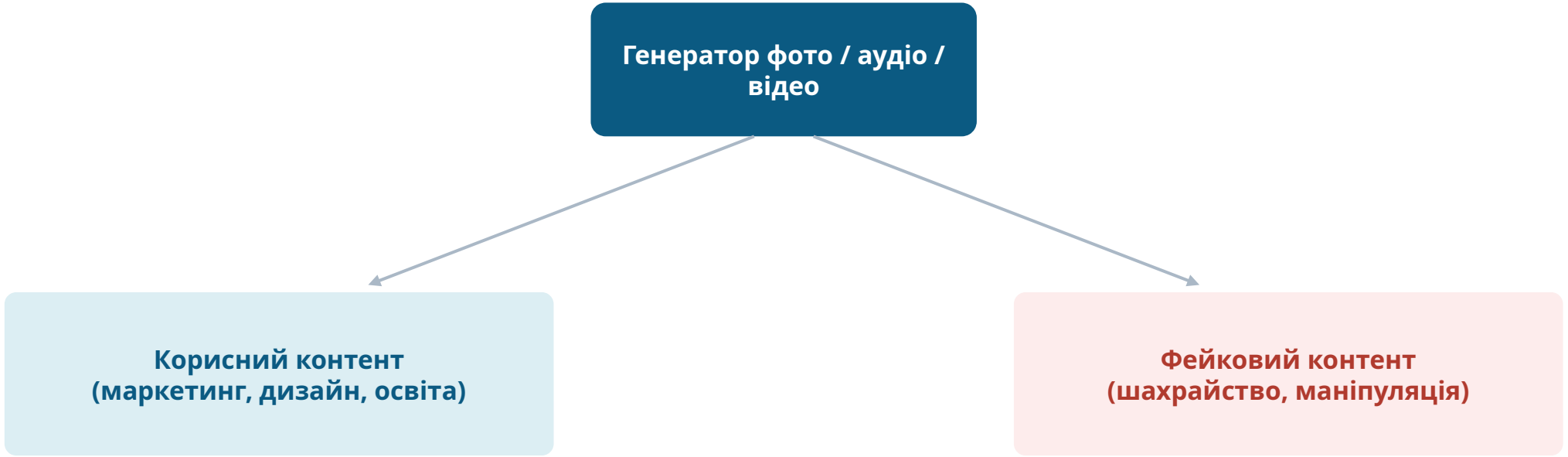
Ефект чорної скриньки



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

ТИП ПОМИЛКИ 5 (НАВМИСНА)

Дипфейки та дезінформація



Та сама технологія генерує і корисний, і шкідливий контент — вона може потрапити в «погані руки» для шахрайства чи маніпуляції громадською думкою.

КЛЮЧОВЕ ПРАВИЛО СПОЖИВАННЯ КОНТЕНТУ

Якщо контент провокує сильну емоцію — перевірте джерело або знайдіть незалежне підтвердження

Емоційно «заряджений» контент — улюблений інструмент і дезінформації, і генеративного ШІ, який оптимізує залучення, а не істину.

ПІДСУМОВУЄМО

Що варто запам'ятати з цієї лекції

- Цільова функція ГШІ — переконливість тексту, а не істина; правильність відповіді — лише побічний ефект
- Галюцинації — не «баг», а наслідок відсутності вбудованого механізму «я не знаю»
- Зростання якості моделей не гарантує зменшення фактичних помилок (приклад — сфабриковані наукові цитування)
- ШІ відтворює упередження навчальних даних і може по-різному висвітлювати дискусійні теми залежно від мови джерел
- Головне правило: впевнений тон ≠ правильна відповідь; емоційний контент — сигнал перевірити джерело

Наступна лекція: промпт-інжиніринг — як формулювати запити, щоб мінімізувати ці ризики.