

ТЕХНОЛОГІЇ ШТУЧНОГО ІНТЕЛЕКТУ ТА БЕЗПЕКА ДАНИХ

Лекція 8

Приватність у діалозі з ШІ: куди йдуть ваші дані



Від даних загалом до конкретного діалогу з ШІ

- Лекція 7 дала загальний понятійний апарат: типи даних, тріаду CIA, основи законодавства
- Ця лекція звужує фокус до найпоширенішої практичної ситуації: діалогу з чат-ботом на кшталт ChatGPT, Claude чи Gemini

Що насправді відбувається з текстом, який ви вводите в чат-бот? Хто його бачить? Куди він потрапляє? І як мінімізувати ризик?

Від параної до безтурботності

1960–70-ті

«Не кажи цього по телефону — можуть підслухати». Побутова параноя щодо прослуховування була широко розповсюдженим явищем.

Сьогодні

Мільйони людей щодня добровільно розмовляють з розумними колонками вдома, ставлять особисті питання, не замислюючись про обробку цих записів.

Психологічний зсув стався не тому, що ризик зник, а тому, що технологія стала звичною — «приватність через звичку», а не через свідому оцінку ризику.

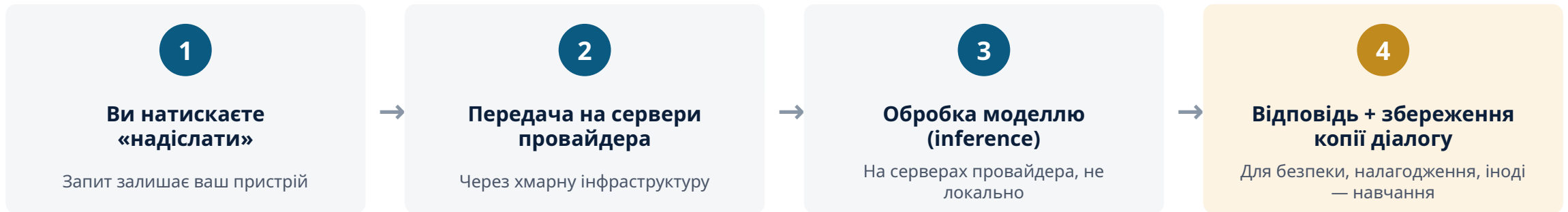
ГОЛОВНИЙ ПРИНЦИП ЛЕКЦІЇ

«Це завжди чийсь комп'ютер»

Усе, що ви вводите в чат-бот — текст, документи, голосові команди — залишає ваш пристрій і обробляється компанією, яка надає сервіс.

Це не привід уникати ШІ — а привід ставитися до нього як до будь-якого іншого онлайн-сервісу, що обробляє особисті дані.

Технічний шлях одного запиту



Між «я написав повідомлення» і «я отримав відповідь» існує кілька точок, де дані потенційно доступні третім сторонам.

- Технічний персонал провайдера
- Системи моніторингу й безпеки
- У певних випадках — процеси донавчання моделі

Архітектура провайдера ШІ



- Провайдер великої мовної моделі — не єдина програма, а ціла інфраструктура з кількох компонентів
- Користувачі надсилають запити (Prompts/Tasks) через застосунки й отримують відповіді

Ваші діалоги можуть ставати матеріалом для донавчання майбутніх версій моделі — технологія з лекції 2 тут стає питанням приватності.

Чотири канали витоку приватності

1 Витік із тренувальних даних

Чи конкретні дані були частиною тренувального набору моделі

2 Витік характеристик моделі

Витягування внутрішніх параметрів через ретельно сформульовані запити

3 Витік з історії розмов

Попередні повідомлення впливають на майбутні відповіді неочікувано

4 Недотримання наміру приватності

Система формально, а не по суті, виконує прохання «не запам'ятовуй»

Ці ризики здебільшого невидимі для звичайного користувача — саме тому важливо знати про них завчасно.

ГОЛОВНЕ ПРАКТИЧНЕ ПРАВИЛО

**Не вводьте в загальнодоступний ШІ те,
що ви не хотіли б, щоб побачив
хтось сторонній**

Аналогія: ви навряд чи стали б вголос диктувати номер банківської картки в переповненому вагоні метро.

Категорії інформації під особливим ризиком



Конфіденційні дані клієнтів і партнерів



Паролі, ключі доступу, токени автентифікації



Медична інформація (своя чи чужа)



Персональні дані інших людей без згоди



Комерційна таємниця й конфіденційна ділова інформація

Публічні versus корпоративні рішення

Публічна версія

Безкоштовний обліковий запис фізичної особи. Політики обробки даних можуть включати використання розмов для тренування — потрібно перевіряти й вимикати вручну.

Корпоративна (enterprise)

Підписка організації з окремою угодою про обробку даних (data processing agreement) — чіткіші гарантії конфіденційності й невикористання для тренування.

Перш ніж використовувати ШІ-інструмент для чутливих даних компанії — з'ясуйте, яка саме версія сервісу використовується.

GDPR у контексті діалогу з ШІ

- GDPR дає людям права на свої дані незалежно від технології, що їх обробляє (пор. права суб'єкта даних з лекції 7)
- **У контексті ШІ це означає:**
 - право знати, чи використовувались ваші дані для тренування моделі
 - право вимагати видалення даних із систем провайдера
 - право на пояснення автоматизованого рішення (напр. у кредитному скорингу)

Технічна складність: «видалити» інформацію з уже натренованої нейронної мережі значно складніше, ніж з рядка бази даних — це сфера досліджень machine unlearning.

EU AI Act: чотири рівні ризику

Неприйнятний — заборонено

Соціальний скоринг громадян державою, маніпулятивні «темні паттерни», розпізнавання облич у реальному часі

Високий

Критична інфраструктура, найм персоналу, освіта, кредитний скоринг, прикордонний контроль, судові системи

Обмежений

Чат-боти та системи з вимогами прозорості — користувач має знати, що спілкується зі ШІ

Мінімальний

Переважає більшість застосунків ШІ — без особливих законодавчих вимог

Перший комплексний закон ЄС про регулювання ШІ (з 2024 року) — класифікує системи за рівнем ризику, а не за типом технології.

Де саме тут звичайний чат-бот?

Більшість чат-ботів загального застосування (ChatGPT, Claude, Gemini) — «обмежений ризик» за AI Act.

- **Ключова вимога — прозорість:**
 - користувач має чітко розуміти, що взаємодіє зі штучним інтелектом, а не з людиною

Важливий нюанс: якщо той самий чат-бот вбудовується в систему прийняття рішень про кредит — вимоги підвищуються до «високого ризику». Рівень ризику визначається застосуванням, а не лише технологією.

Правила безпечного діалогу з ШІ

- ✓ Застосуйте чек-лист класифікації даних з лекції 7 перш ніж вводити інформацію
- ✓ Використовуйте enterprise-рішення для робочих задач із чутливими даними
- ✓ Перевіряйте й вмикайте налаштування «не використовувати для тренування»
- ✓ Анонімізуйте дані, де можливо (напр. «Клієнт А» замість реального імені)
- ✓ Пам'ятайте: видалити повідомлення з чату ≠ видалити дані з усіх систем провайдера

Як це пов'язано з попередніми лекціями

Лекція 2 · RLHF

Та сама петля зворотного зв'язку потенційно перетворює ваші розмови на навчальний матеріал

Лекція 4 · Чорна скринька

Не можемо простежити, як конкретний фрагмент введеної інформації вплинув на систему в довгостроковій перспективі

Лекція 7 · Дані як актив

Ваші розмови з ШІ — це так само актив, тепер уже для провайдера сервісу

ПІДСУМОВУЄМО

Що варто запам'ятати з цієї лекції

- «Це завжди чийсь комп'ютер» — усе введене в чат-бот залишає ваш пристрій і обробляється провайдером
- Існує чотири специфічні канали витоку приватності в діалозі зі ШІ, здебільшого невидимі для користувача
- Головне правило: не вводьте те, що не хотіли б показати сторонньому
- GDPR дає права на свої дані, але «видалення» з натренованої моделі технічно складне
- EU AI Act класифікує системи за рівнем ризику — звичайний чат-бот належить до «обмеженого», але застосування може підвищити цей рівень

Наступна лекція: право і регулювання ШІ — авторське право на згенерований контент і відповідальність за рішення.