

Технології штучного інтелекту та безпека даних

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

Лекція 8. Приватність у діалозі з ШІ: куди йдуть ваші дані

Мета лекції: прослідкувати технічний і юридичний шлях звичайного запиту до чат-бота — від натискання «надіслати» до можливого використання цих даних у майбутньому — і дати студентам конкретні практичні правила безпечного діалогу з генеративним ШІ.

Зв'язок з попередніми лекціями: лекція 7 дала загальний понятійний апарат — типи даних, тріаду CIA, основи законодавства. Ця лекція звужує фокус до найпоширенішої практичної ситуації: діалогу з чат-ботом, — і прямо застосовує до неї RLHF (лекція 2) та ефект чорної скриньки (лекція 4).

1. Вступ: від даних загалом до конкретного діалогу з ШІ

- Лекція 7 дала загальний понятійний апарат: що таке дані, як їх класифікувати, які основні юридичні поняття існують.
- Ця лекція звужує фокус до найпоширенішої практичної ситуації для студента чи працівника: звичайного діалогу з чат-ботом на кшталт ChatGPT, Claude чи Gemini.
- Питання лекції: що насправді відбувається з текстом, який ви вводите в чат-бот? Хто його бачить? Куди він потрапляє? І як мінімізувати ризик випадкового розкриття чутливої інформації?

2. Історична паралель: від параної до безтурботності

- У 1960–70-х роках побутова параноя щодо прослуховування телефонних розмов була широко розповсюдженим явищем — «не кажи цього по телефону, можуть підслухати».
- Сьогодні мільйони людей щодня добровільно розмовляють з розумними колонками та голосовими асистентами вдома, ставлять їм особисті питання, не замислюючись, що ці записи так само обробляються сторонньою компанією.
- Психологічний зсув стався не тому, що ризик зник, а тому, що технологія стала звичною і зручною — **«приватність через звичку»** (habituation), а не через свідому оцінку ризику.

Урок для курсу: зручність технології не повинна автоматично означати відсутність ризику — варто зберігати свідомий, а не автоматичний підхід до того, чим ділишся з цифровими системами.

3. Головний принцип: «це завжди чийсь комп'ютер»

- Популярний вислів «there is no cloud, it's just someone else's computer» влучно передає суть: коли ви користуєтесь хмарним сервісом (включно з чат-ботом), ваші дані фізично обробляються й зберігаються на серверах, що належать іншій компанії.

- 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 (як і не привід відмовлятися від електронної пошти чи хмарного сховища), а привід ставитися до нього як до будь-якого іншого онлайн-сервісу, що обробляє особисті дані — з відповідною обережністю.

4. Технічний шлях одного запиту

- Щоб зрозуміти, де саме виникають ризики приватності, корисно простежити шлях звичайного запиту від моменту, коли ви натискаєте «надіслати».
- **Крок 1:** запит залишає ваш пристрій і передається на сервери провайдера ШІ-сервісу (часто через посередників — хмарну інфраструктуру).
- **Крок 2:** запит обробляється моделлю (inference) — обчислювально інтенсивний процес, що відбувається на серверах провайдера, а не локально на вашому пристрої.
- **Крок 3:** відповідь генерується і повертається користувачу; водночас копія діалогу зазвичай зберігається провайдером (для безпеки, налагодження, іноді — для подальшого навчання моделей, залежно від налаштувань і типу підписки).

Висновок: між «я написав повідомлення» і «я отримав відповідь» існує кілька точок, де дані потенційно доступні третім сторонам — технічному персоналу провайдера, системам моніторингу, а в певних випадках — процесам донавчання моделі.

5. Архітектура провайдера ШІ: від тренувальних даних до вашого запиту

- Провайдер великої мовної моделі (LLM Provider) — це не єдина програма, а ціла інфраструктура з кількох компонентів.
- **Тренувальний набір даних** (публічний і власницький) використовується для попереднього тренування базової моделі (Pre-trained LLM).
- Модель постачається через програмний інтерфейс (**LLM API**), до якого підключаються кінцеві застосунки — вебчат, мобільний застосунок, інтеграції в інші програми.
- Користувачі надсилають запити (**Prompts/Tasks**) через ці застосунки, отримують відповіді, і — що важливо — цей обмін часто формує «петлю зворотного зв'язку» (**Feedback loop**).

Дидактична примітка: Ця петля зворотного зв'язку — той самий механізм RLHF з лекції 2: ваші діалоги можуть ставати матеріалом для донавчання майбутніх версій моделі. Технологічне пояснення з лекції 2 тут [прямо перетворюється на юридично-етичне питання приватності](#).

6. Чотири канали витоку приватності в діалозі з ШІ

- Дослідники конфіденційності генеративного ШІ виділяють кілька специфічних каналів, якими персональна чи чутлива інформація може «просочуватися» через систему.

Канал	Що відбувається
Витік членства й властивостей з тренувальних даних	За певних умов можна визначити, чи конкретні дані (напр. текст конкретної людини) були частиною тренувального набору моделі
Витік характеристик моделі	Витягування внутрішніх параметрів чи закономірностей навченої моделі через ретельно сформульовані запити
Неявний витік з історії розмов	Інформація з попередніх повідомлень у межах сесії чи облікового запису може вплинути на майбутні відповіді неочікуваним чином
Дотримання наміру приватності користувача	Наскільки система справді виконує прохання користувача щодо приватності (напр. «не запам'ятовуй це»), а не лише формально погоджується

Практичний висновок: ці ризики здебільшого невидимі для звичайного користувача під час діалогу — саме тому важливо знати про них і поводитися обережно завчасно, а не постфактум.

7. Мінімізація чутливої інформації: головне практичне правило

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

- Це правило узагальнює всю технічну складність попередніх пунктів в одну просту евристику, застосовну без глибокого розуміння архітектури моделі.
- Аналогія: так само, як ви навряд чи стали б вголос диктувати номер банківської картки в переповненому вагоні метро, не варто вводити його в чат-бот «про всяк випадок».

8. Категорії інформації під особливим ризиком

Пряме продовження класифікації даних з лекції 7 — застосоване саме до контексту введення в ШІ:

- **Конфіденційні дані клієнтів і партнерів** — імена, контакти, історія замовлень, договірні умови.
- **Паролі, ключі доступу, токени автентифікації** — для «перевірки» чи «форматування» ніколи не варто вводити реальні креденшили в чат-бот.
- **Медична інформація** — як власна, так і про інших людей, зокрема пацієнтів (для тих, хто працює в медичній чи суміжній сфері).
- **Персональні дані інших людей без їхньої згоди** — навіть якщо мета видається невинною (наприклад, «напиши характеристику на колегу»).
- **Комерційна таємниця й конфіденційна ділова інформація компанії** — фінансові показники до публікації, стратегічні документи, вихідний код власницьких систем.

9. Публічні versus корпоративні рішення: чи є різниця?

- Безкоштовні, загальнодоступні версії чат-ботів для фізичних осіб і корпоративні (enterprise) підписки для організацій часто мають різні політики обробки даних.
- Багато провайдерів пропонують налаштування «не використовувати мої розмови для тренування моделі» — варто перевіряти й свідомо вмикати такі опції, де це можливо.
- Корпоративні угоди (enterprise agreements, data processing agreements) зазвичай містять чіткіші гарантії щодо конфіденційності, зберігання і невикористання даних клієнта для тренування — але це предмет окремого договору, а не автоматична гарантія.

Практичний висновок для майбутнього фахівця: перш ніж використовувати ШІ-інструмент для роботи з чутливими даними компанії, варто з'ясувати, яка саме версія сервісу використовується і які умови обробки даних вона передбачає.

10. GDPR у контексті діалогу з ШІ

- Загальний регламент про захист даних (GDPR) Європейського Союзу дає людям права на свої дані незалежно від того, яка технологія їх обробляє (пор. права суб'єкта даних з лекції 7).
- У контексті ШІ це означає, зокрема: право знати, чи використовувалися ваші персональні дані для тренування моделі; право вимагати видалення своїх даних із систем провайдера (наскільки це технічно можливо); право на пояснення автоматизованого рішення, якщо ШІ-система прямо впливає на вас (наприклад, у кредитному скорингу).

Технічна складність: на відміну від традиційної бази даних, «видалити» конкретну інформацію з уже натренованої нейронної мережі технічно значно складніше, ніж з рядка таблиці — це активна сфера досліджень, відома як machine unlearning.

11. EU AI Act: регулювання за рівнем ризику

- Перший комплексний закон ЄС про регулювання самого штучного інтелекту (набув чинності поетапно з 2024 року) класифікує ШІ-системи за рівнем ризику, а не за типом технології.

Рівень ризику	Приклади
Неприйнятний (заборонено)	Соціальний скоринг громадян державою, маніпулятивні «темні паттерни» ШІ, певні види розпізнавання облич у реальному часі в публічних місцях
Високий	Критична інфраструктура, найм і оцінка персоналу, доступ до освіти й публічних послуг, кредитний скоринг, прикордонний контроль, судові системи
Обмежений	Системи із специфічними вимогами прозорості, включно з чат-ботами — користувач повинен знати, що спілкується зі ШІ
Мінімальний	Переважна більшість застосунків ШІ (відеоігри зі ШІ, спам-фільтри) — без особливих законодавчих вимог

12. Де саме тут звичайний чат-бот?

- Більшість чат-ботів для загального застосування (ChatGPT, Claude, Gemini тощо) підпадають під категорію «обмеженого ризику» за AI Act.
- Ключова вимога для цієї категорії — **прозорість**: користувач має чітко розуміти, що взаємодіє зі штучним інтелектом, а не з людиною.

Важливий нюанс: якщо той самий чат-бот вбудовується в систему з вищим рівнем ризику (наприклад, як інтерфейс для автоматизованого прийняття рішень про кредит), то вимоги до всієї системи підвищуються до рівня «високого ризику» — тобто **рівень ризику визначається застосуванням, а не лише базовою технологією**.

13. Практичні правила безпечного діалогу з ШІ

- Перед тим як ввести інформацію в чат-бот, подумки застосуйте чек-лист класифікації даних з лекції 7.
- Використовуйте корпоративні/enterprise рішення для робочих задач із чутливими даними, а не особисті облікові записи в публічних сервісах.
- Перевіряйте й свідомо налаштовуйте параметри конфіденційності (можливість вимкнути використання діалогів для тренування моделі).
- Анонімізуйте чи узагальнюйте дані перед введенням, де це можливо (наприклад, замініть реальні імена клієнтів на «Клієнт А», «Клієнт Б»).
- Пам'ятайте: видалити повідомлення з інтерфейсу чату — не те саме, що гарантовано видалити дані з усіх систем провайдера.

14. Наскрізні зв'язки з попередніми лекціями

- **RLHF (лекція 2):** та сама петля зворотного зв'язку, що покращує моделі, потенційно перетворює ваші розмови на навчальний матеріал для майбутніх версій.
- **Ефект чорної скриньки (лекція 4):** якщо ми не можемо повністю пояснити, чому модель дала конкретну відповідь, ми так само не можемо гарантовано простежити, як саме конкретний фрагмент введеної інформації вплинув на систему в довгостроковій перспективі.
- **Дані як економічний актив (лекція 7):** ваші розмови з ШІ — це так само актив, тепер уже для провайдера сервісу, а не тільки для вас чи вашої компанії.

Місток до наступної лекції: ми розглянули, куди насправді йдуть дані в діалозі з ШІ, які канали витоку приватності існують, і як GDPR та AI Act регулюють цю сферу з погляду прав людини й рівня ризику. Наступна лекція розширює тему регулювання: детальніше розберемо авторське право на згенерований контент і відповідальність за рішення, ухвалені на основі відповіді ШІ. Це тема лекції 9.

Контрольні питання для самоперевірки студентів

- Поясніть своїми словами вислів «це завжди чийсь комп'ютер» і чому він не є аргументом проти використання ШІ.
- Опишіть шлях одного запиту від натискання «надіслати» до отримання відповіді — які точки потенційного ризику ви можете назвати?
- Наведіть приклад ситуації з власного досвіду (навчання, підробіток), коли ви мало не ввели чи ввели в чат-бот інформацію, яку варто було б класифікувати як чутливу.
- Чим публічна версія чат-бота відрізняється від корпоративної з погляду обробки даних?
- До якої категорії ризику за EU AI Act належить звичайний чат-бот загального призначення? А якщо той самий чат-бот використовується для автоматизованого відбору кандидатів на роботу?
- Чому «видалити» дані з уже натренованої моделі технічно складніше, ніж з таблиці бази даних?