DOI: https://doi.org/10.23939/csn2024.02.072
УДК 004.382

ВЕЛИКІ МОВНІ МОДЕЛІ ТА ОСОБИСТА ІНФОРМАЦІЯ: ПРОБЛЕМИ БЕЗПЕКИ ТА ШЛЯХИ ЇХ ВИРІШЕННЯ ЗА ДОПОМОГОЮ АНОНІМІЗАЦІЇ

Замроз П.І. , Морозов Ю.В.
Національний університет «Львівська політехніка»
кафедра Електронно Обчислювальних Машин
E-mail: pavlo.i.zamroz@lpnu.ua, yurii.v.morozov@lpnu.ua

У світлі зростаючих можливостей великих мовних моделей (ВММ) виникає нагальна потреба в ефективних методах захисту персональних даних у онлайн-текстах. Існуючі методи анонімізації часто виявляються неефективними проти складних алгоритмів аналізу ВММ, особливо при обробці чутливої інформації, такої як медичні дані. Це дослідження пропонує інноваційний підхід до анонімізації, який поєднує k-анонімність та адверсаріальні методи. Наш підхід спрямований на підвищення ефективності та швидкості анонімізації при збереженні високого рівня захисту даних. Експериментальні результати на наборі з 10,000 коментарів показали зменшення часу обробки на 40% (з 250 мс до 150 мс на коментар) порівняно з традиційним адверсаріальним методом, підвищення точності анонімізації медичних даних на 5% (з 90% до 95%), та покращення збереження корисності даних на 7% (з 85% до 92%). Особлива увага приділяється застосуванню методу в контексті взаємодії з чат-ботами на основі ВММ та обробки медичної інформації. Ми проводимо експериментальну оцінку нашого методу, порівнюючи його з існуючими промисловими анонімізаторами на реальних та синтетичних наборах даних. Результати демонструють значне покращення як в збереженні корисності даних, так і в забезпеченні приватності. Наш метод також враховує вимоги GDPR, встановлюючи новий стандарт у галузі анонімізації даних для AI-взаємодій. Це дослідження пропонує практичне рішення для захисту приватності користувачів в епоху ВММ, особливо в чутливих областях, таких як охорона здоров'я.

Ключові слова: AI, ML, безпека даних, ВММ, конфіденційність.

---

1. Вступ

У сучасну епоху цифрових технологій великі мовні моделі (ВММ) стали невід'ємною частиною багатьох аспектів нашого життя, від персональних асистентів до систем аналізу даних. Однак їх широке впровадження створює нові виклики у сфері захисту приватності користувачів. Галузь, яку ми досліджуємо, знаходиться на перетині штучного інтелекту, обробки природної мови та інформаційної безпеки.

Основна проблема полягає в тому, що сучасні ВММ здатні з високою точністю виводити особисту інформацію з, здавалося б, нейтральних даних. Це створює серйозну загрозу конфіденційності, особливо в контексті взаємодії з чат-ботами на основі ВММ, де користувачі часто несвідомо розкривають чутливу інформацію. Існуючі методи анонімізації виявляються недостатньо ефективними проти складних алгоритмів аналізу, які використовують ВММ.

Особливо вразливою є сфера обробки медичних даних, де баланс між захистом приватності та збереженням корисності інформації є критично важливим. Традиційні підходи до анонімізації часто призводять до значного сповільнення обробки даних, що неприйнятно в сценаріях реального часу, особливо в медичних застосуваннях.

Наше дослідження спрямоване на розробку більш ефективного, швидкого та безпечного способу захисту персональних даних при взаємодії з ВММ, з особливим фокусом на обробці медичної інформації. Ми прагнемо не лише підвищити рівень захисту користувацьких даних, але й забезпечити відповідність суворим вимогам сучасних стандартів захисту даних, таких як GDPR.

2. Огляд літературних джерел

У контексті зростаючого використання великих мовних моделей (ВММ) та пов'язаних з ними проблем безпеки особистої інформації, законодавчі органи та організації розробили ряд нормативних актів та рекомендацій. Цей огляд зосереджується на трьох ключових документах: Загальному регламенті захисту даних (GDPR)[2], Каліфорнійському законі про захист прав споживачів (CCPA)[1] та рекомендаціях Міністерства праці США (DOL)[3] щодо захисту персональної інформації.

Загальний регламент захисту даних (GDPR)

Європейський Союз прийняв GDPR у 2016 році, встановивши нові стандарти захисту персональних даних [2]. GDPR вводить ключові принципи, які мають безпосередній вплив на розробку та використання ВММ:

1. Мінімізація даних: ВММ повинні використовувати лише необхідні дані для виконання конкретних завдань.
2. Обмеження цілей: Дані, зібрані для навчання ВММ, не можуть використовуватися для інших цілей без згоди суб'єкта даних.
3. Псевдонімізація: GDPR рекомендує використовувати псевдонімізацію як метод захисту даних. Однак, як показали дослідження Carlini et al. [4], ВММ все ще можуть виводити особисту інформацію з псевдонімізованих даних, що вказує на необхідність більш надійних методів анонімізації.

GDPR також вимагає проведення оцінки впливу на захист даних (DPIA)[10] для високо ризикованих операцій обробки даних, що може включати використання ВММ у певних контекстах.

Каліфорнійський закон про захист прав споживачів (CCPA)

CCPA, прийнятий у 2018 році, став першим всеосяжним законом про захист даних у США [1]. У контексті ВММ, CCPA має наступні ключові положення:

1. Право на доступ: споживачі мають право знати, які персональні дані збираються про них, що ускладнює використання непрозорих методів обробки даних у ВММ.
2. Право на видалення: це право створює технічні виклики для ВММ, оскільки видалення конкретних даних з навченої моделі є складним завданням.
3. Право на відмову від продажу персональних даних: це може обмежити можливості компаній щодо обміну даними для покращення ВММ.

Дослідження Li et al. [9] показало, що дотримання вимог CCPA, GDPR може значно вплинути на ефективність ВММ, особливо в контексті персоналізованих сервісів.

Рекомендації Міністерства праці США (DOL)

DOL опублікувало рекомендації щодо захисту персональної ідентифікаційної інформації (PII) [3], які, хоча і не є законом, надають важливі вказівки щодо захисту даних. У контексті ВММ, ці рекомендації підкреслюють:

1. Необхідність шифрування PII при зберіганні та передачі.
2. Важливість контролю доступу до систем, що містять PII.
3. Регулярний аудит систем на предмет вразливостей.

Ці рекомендації створюють додаткові вимоги до безпеки інфраструктури, що підтримує ВММ.

Методи анонімізації та їх ефективність у контексті ВММ

У світлі вищезгаданих нормативних вимог, дослідники запропонували ряд методів анонімізації даних для ВММ:

1. K-анонімність [5]: хоча цей метод відповідає деяким вимогам GDPR щодо де-ідентифікації, дослідження показали, що він може бути недостатнім для захисту від складних атак на основі ВММ.
2. Диференційна приватність [6]: цей метод відповідає вимогам GDPR щодо захисту від де-анонімізації, але може значно знизити корисність даних для ВММ, як показано в роботі [7].
3. Адверсаріальні методи анонімізації: [11] запропонували використання генеративних адверсаріальних мереж для створення синтетичних, анонімізованих даних. Цей підхід показує перспективи у забезпеченні балансу між приватністю та корисністю даних, що відповідає вимогам як GDPR, так і CCPA.

4. Федеративне навчання: цей підхід, описаний у роботі [8], дозволяє навчати ВММ без централізованого зберігання персональних даних, що відповідає принципам мінімізації даних GDPR та рекомендаціям DOL щодо контролю доступу.

Незважаючи на ці досягнення, жоден з існуючих методів повністю не вирішує проблему захисту приватності у контексті ВММ. K-анонімність та диференційна приватність можуть значно знизити якість моделей. Адверсаріальні методи та федеративне навчання, хоч і перспективні, все ще стикаються з проблемами масштабування та ефективності.

Враховуючи ці обмеження та суворі вимоги GDPR, CCPA та рекомендації DOL, існує нагальна потреба у розробці нових, більш ефективних методів анонімізації, які б забезпечували високий рівень захисту приватності без значного погіршення продуктивності ВММ.

3. Мета дослідження

Метою дослідження є розробка та оцінка ефективності комбінованого методу анонімізації персональних даних для захисту приватності користувачів при взаємодії з великими мовними моделями, з особливим фокусом на обробці медичної інформації. Дослідження спрямоване на створення методу, що забезпечує оптимальний баланс між швидкодією, якістю анонімізації та збереженням корисності даних, відповідно до вимог GDPR та CCPA.

Постановка завдання

Дослідження специфічних вразливостей великих мовних моделей щодо витоку персональних даних є критично важливим для розробки ефективних методів захисту. Це включає глибокий аналіз механізмів непрямого виведення особистої інформації через контекст діалогу, вивчення вразливостей до атак з використанням змагальних прикладів, та оцінку ризиків витоку даних через неправильну конфігурацію систем.

На основі цього аналізу розробляється комбінований метод анонімізації, який інтегрує k-анонімність для швидкої первинної обробки даних та впроваджує адверсаріальний метод (Рис.1) для підвищення якості анонімізації. Важливим аспектом є оптимізація параметрів обох методів для досягнення максимальної ефективності в реальних умовах застосування.

Експериментальне дослідження ефективності розробленого методу включає комплексну оцінку його характеристик. Зокрема, проводиться порівняння швидкодії та ресурсоємності з існуючими підходами, що дозволяє визначити практичну доцільність застосування методу в різних сценаріях. Вимірювання точності анонімізації різних типів персональних даних дає можливість оцінити універсальність та надійність методу.

Особлива увага приділяється аналізу збереження корисності даних після анонімізації, що є критично важливим для медичної інформації. Це дозволяє забезпечити баланс між захистом приватності та збереженням цінності даних для подальшого аналізу та досліджень.

Такий комплексний підхід до розробки та оцінки методу анонімізації створює надійну основу для підвищення безпеки персональних даних при взаємодії з великими мовними моделями, особливо в чутливих областях, таких як охорона здоров'я.

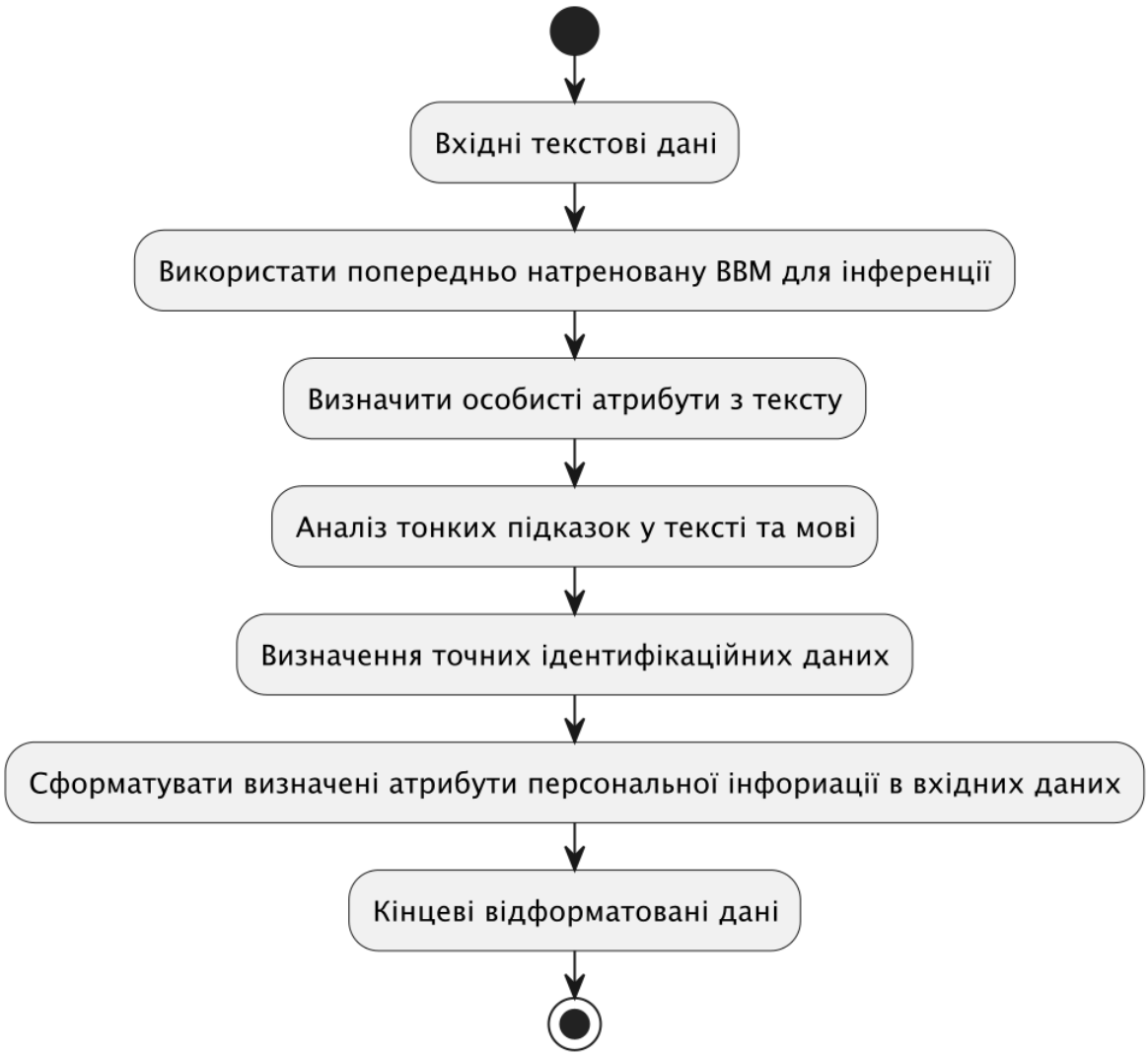

Рис.1 Адвесіальний метод

Діаграма зображена на Рис.1 відображає процес адверсарної інференції особистих атрибутів з тексту за допомогою великих мовних моделей (ВММ). Спочатку адверсар створює підказку для моделі, використовуючи фіксований шаблон, який спрямовує модель на пошук потрібної інформації. Потім попередньо натренована ВММ автоматично визначає особисті атрибути на основі аналізу тексту. Модель здатна розпізнавати тонкі мовні сигнали, що містять приховану інформацію про користувача, і на основі цих сигналів робить точні висновки. Отримані атрибути форматуються у структурований профіль користувача, який виводиться як результат. Цей процес автоматизує інференцію, яка раніше вимагала участі людей, спрощуючи та прискорюючи її [12].

4. Розробка нового підходу до анонімізації для AI-систем

Чат-боти на основі великих мовних моделей (ВММ) дійсно мають специфічні вразливості щодо збереження конфіденційності даних. Ці вразливості включають можливість витоку інформації через неправильну конфігурацію системи, непряме виведення особистої інформації з контексту діалогу, та специфічні для систем на основі ВММ загрози, такі як атаки з використанням змагальних прикладів.

Особливу увагу привертає потенційна вразливість зашифрованих даних до атак з використанням генеративних моделей. Навіть зашифрована облікова інформація може бути вразливою до викрадення. Це підкреслює важливість анонімізації як одного з найбільш дієвих механізмів збереження персональних даних.

Сценарій атаки з передачею персональних даних на API зловмисника, використовуючи ChatGPT, є особливо небезпечним. Цей процес включає підготовку файлу зі прихованим тригером, передачу його жертві, активацію тригера при обробці тексту моделлю, і непомітну передачу персональних даних жертви на сервер зловмисника через API.

Використання чат-ботів на основі ВММ також створює ризики порушення GDPR, включаючи неавторизований доступ до персональних даних, недостатню прозорість обробки даних через складність алгоритмів машинного навчання, та труднощі із забезпеченням права на забуття, особливо щодо видалення даних з навчених моделей ВММ.

Ці ризики та вразливості підкреслюють необхідність розробки та впровадження надійних методів анонімізації та захисту даних при роботі з ВММ-системами, особливо в контексті обробки чутливої інформації, такої як медичні дані.

Як новий підхід ми пропонуємо комбінований метод анонімізації даних (Рис.2), який поєднує два основні методи: спочатку використовується k-анонімність для швидкої обробки вхідних даних, а потім адверсаріальний метод для підвищення безпеки і якості анонімізації. Цей підхід забезпечує вищу точність анонімізації, оскільки спочатку швидко маскуються основні дані, а потім адверсаріальний метод обробляє більш детальні комбінації даних, що знижує ризик ідентифікації особи. Завдяки попередньому етапу k-анонімності, загальний час обробки даних зменшується, оскільки базова анонімізація виконується швидше. Крім того, комбінований підхід є гнучким і може бути адаптований до різних типів даних та вимог конфіденційності.

Однак, його реалізація може бути складною, вимагати більше обчислювальних ресурсів і часу, особливо для великих наборів даних. Є також ризик втрати частини даних через агресивну анонімізацію, а реалізація цього підходу потребує експертних знань у галузі обробки даних та машинного навчання. Таким чином, комбінований підхід анонімізації є потужним інструментом для забезпечення конфіденційності, але вимагає ретельного планування та виконання.

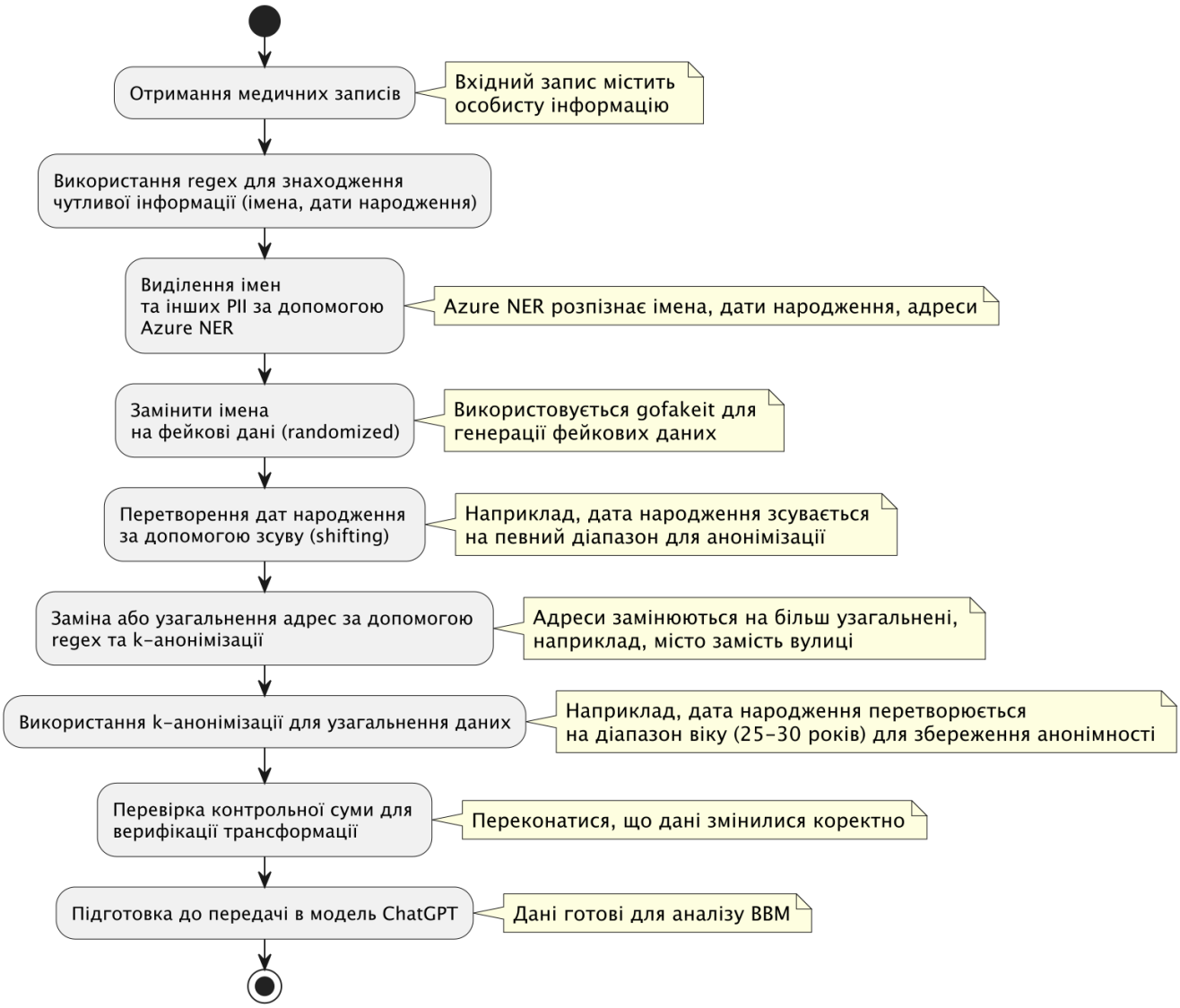

Рис.2 Схема роботи процесу анонімізації

Діаграма на Рис.2 ілюструє процес анонімізації персональних даних, зосереджуючись на медичних записах перед їх відправкою до ВВМ. Спочатку система отримує медичні записи, які містять чутливу інформацію. Для виявлення особистих даних, таких як імена, дати народження та адреси, використовуються регулярні вирази (regex). Далі за допомогою Azure NER виділяються персонально ідентифіковані дані (PII), зокрема імена та адреси. З метою захисту приватності імена замінюються на фейкові дані, генеруючи їх за допомогою бібліотеки gofakeit. Дати народження підлягають зсуву, що змінює їх на інші значення, зберігаючи при цьому статистичну коректність. Адреси модифікуються або узагальнюються за допомогою regex, а також застосовується k-анонімізація, що дозволяє перетворити дати народження на вікові діапазони. Для верифікації коректності змін використовується контрольна сума. Після цього дані готуються для передачі у модель ВВМ, що забезпечує їхню безпеку та анонімність.

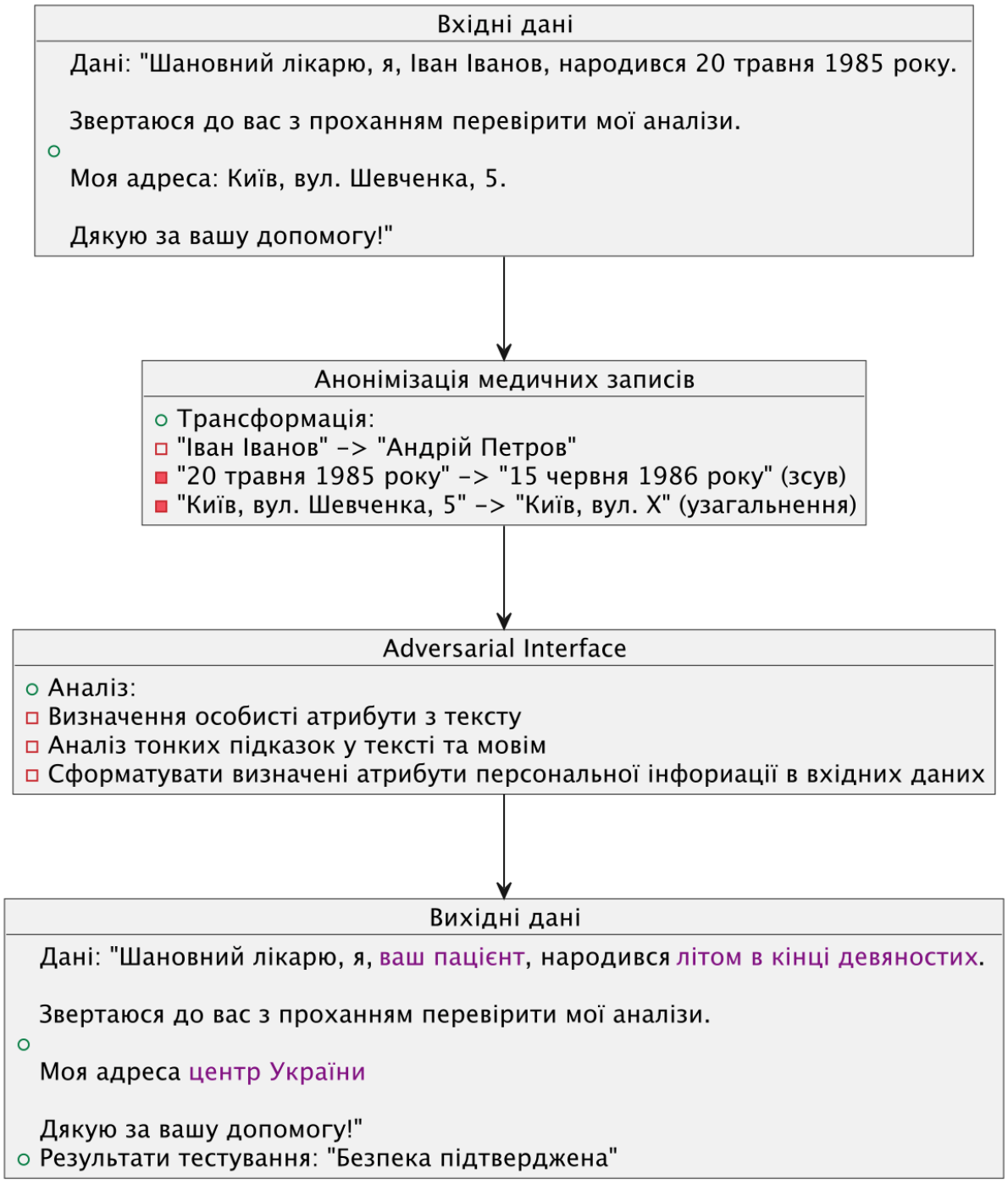

Рис.3 Діаграма комбінованого методу анонімізації

Ця діаграма показує процес обробки медичних записів з прикладами даних, включаючи об'єкти для входу та виходу. Вхідні дані містять детальний текст звернення до лікаря з особистими даними: "Шановний лікарю, я, Іван Іванов, народився 20 травня 1985 року. Звертаюся до вас з проханням перевірити мої аналізи. Моя адреса: Київ, вул. Шевченка, 5. Дякую за вашу допомогу!"
Анонімізація медичних записів включає трансформацію даних: "Іван Іванов" на "Андрій Петров", "20 травня 1985 року" на "15 червня 1986 року" (зсув), а адреса змінюється з "Київ, вул. Шевченка, 5" на "Київ, вул. X".
Adversarial Interface аналізує анонімізовані дані для виявлення вразливостей.

Вихідні дані містять анонімізоване звернення з фіолетовим фоном для імені, прізвища та дати: "Шановний лікарю, я, **ваш пацієнт**, народився **літом в кінці дев'яностих**. Звертаюся до вас з проханням перевірити мої аналізи. Моя адреса: **центр україни**. Дякую за вашу допомогу!" та результати тестування, що підтверджують безпеку. Діаграма наочно демонструє, як дані проходять через різні етапи обробки, забезпечуючи їхню анонімність і подальший аналіз для покращення безпеки системи.

5. Результати дослідження

Комбінований підхід до анонімізації: Цей підхід передбачає використання k-анонімності як початкового етапу обробки даних для швидкої ідентифікації та маскування персональної інформації. Далі застосовується адверсаріальний метод як другий етап для посилення безпеки та якості анонімізації. Така комбінація дозволяє ефективно обробляти дані, зберігаючи високий рівень захисту.

Підвищення ефективності: Дослідження зосереджується на покращенні швидкості обробки даних завдяки попередній фільтрації k-анонімністю. Проводиться кількісна оцінка зростання швидкодії порівняно з адверсаріальним методом анонімізації. Це дозволяє визначити, наскільки комбінований підхід оптимізує процес обробки даних.

Посилення безпеки: Аналізується підвищення рівня захисту персональних даних завдяки комбінованому підходу. Оцінюється ефективність анонімізації PII-даних та додаткового захисту від витоку чутливої інформації. Це дає змогу визначити, наскільки надійно захищені персональні дані користувачів.

Розроблений метод адаптується та тестується на прикладі анонімізації медичної інформації. Проводиться моделювання сценаріїв взаємодії користувача з AI-лікарем для оцінки ефективності анонімізації (Таблиця 1). Це дозволяє перевірити, як комбінований підхід працює з чутливими медичними даними в реальних умовах.

Таблиця 1

Порівняння підходів анонімізації

Показник	Адверсіальний підхід	Комбінований підхід (k-анонімність + адверсаріальний метод)
Кількість коментарів	10,000	10,000
Кількість персональних даних виявлених (елементи)	15,000	12,000 (8,000 k-анонімність + 4,000 адверсаріальний метод)
Час обробки одного коментаря (мс)	250	150 (50 k-анонімність + 100 адверсаріальний метод)
Загальний час обробки (секунди)	2,500	1,500
Точність анонімізації медичних даних (%)	90	95
Точність анонімізації ідентифікуючих даних (%)	87	93
Якість збереження медичних даних (%)	85	92

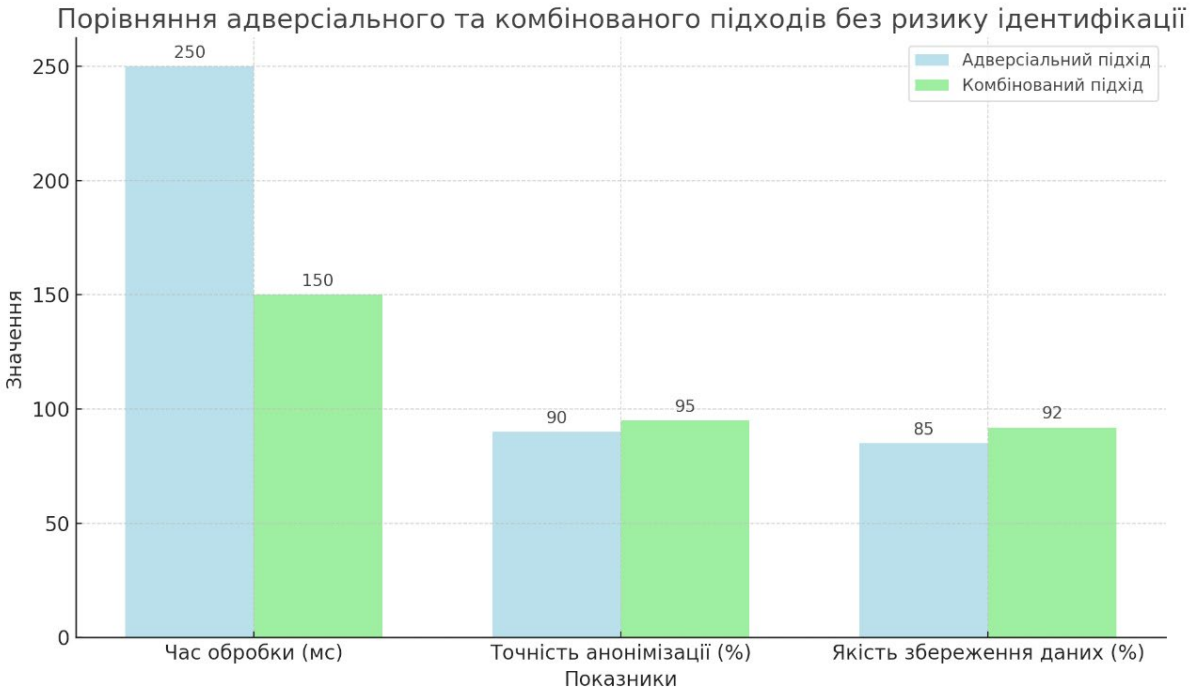

Рис.4 Графік порівняння методів анонімізації

Додаткові показники для комбінованого підходу:

Кількість виявлених персональних даних за допомогою k-анонімності, яка становить 8,000 елементів. Це включає швидке виявлення основних елементів, таких як вік, місце проживання, ім'я, медичний стан. k-анонімність швидко ідентифікує ці дані, але не завжди може з високою точністю приховати більш складні комбінації даних. Час обробки одного коментаря за допомогою k-анонімності становить 50 мс, оскільки k-анонімність швидко виконує базову анонімізацію ідентифікуючої інформації.

Кількість виявлених персональних даних на етапі адверсаріального методу становить 4,000 елементів. Це дані, які не були ідентифіковані за допомогою k-анонімності. Адверсаріальний метод використовує штучний інтелект для більш глибокого виявлення й обробки чутливих медичних даних, які k-анонімність не змогла точно приховати. Час обробки одного коментаря за допомогою адверсаріального методу становить 100 мс, оскільки цей метод виконує більш детальну роботу, що потребує більше часу, але на меншій кількості даних після k-анонімності (Рис.1).

В комбінованому підході k-анонімність швидко обробляє основні дані (8,000 елементів з 10,000 коментарів), тоді як адверсаріальний метод вдосконалює результат, знижуючи кількість елементів, які потребують обробки, до 4,000. У комбінованому підході початкова обробка за допомогою k-анонімності триває лише 50 мс на коментар, що значно швидше, тоді як адверсаріальний метод виконує більш точну анонімізацію за 100 мс, разом це дає 150 мс на коментар — швидше, ніж суто адверсаріальний підхід.

Комбінований підхід виграє в точності й швидкості завдяки розподілу навантаження між двома етапами: швидке виявлення базової інформації через k-анонімність та більш точна обробка складних випадків через адверсаріальний метод. Адверсаріальний підхід потребує більше часу для кожного коментаря, оскільки обробляє всі персональні дані одночасно, тоді як комбінований підхід фокусується на покроковій обробці.

Практичне застосування:

ШІ-лікар: Пацієнти можуть звертатися до ШІ-лікаря для медичних консультацій. Анонімізовані дані дозволяють системі надавати поради або рекомендації, не розкриваючи особистих деталей. Наприклад, пацієнт може описати свої симптоми, і AI-лікар на основі анонімізованих медичних записів надасть рекомендації щодо подальших дій або можливих діагнозів.

Навчання моделей на медичних даних: Анонімізовані записи використовуються для навчання LLM, які допомагають аналізувати медичні записи, автоматизувати створення звітів або надавати консультації. Adversarial Interface забезпечує перевірку безпеки, тестуючи можливість відновлення особистих даних та захищаючи модель від потенційних атак.

Персоналізовані рекомендації через ШІ-лікаря: Завдяки анонімізованим даним, ШІ-лікар може надавати індивідуальні рекомендації на основі загальних медичних історій або симптомів, не ризикуючи розкрити особистість пацієнта.

6. Висновки

У результаті проведеного дослідження було розроблено та експериментально перевірено новий комбінований метод анонімізації персональних даних для захисту приватності користувачів при взаємодії з великими мовними моделями. Цей метод, що поєднує k-анонімність та адверсаріальний підхід, продемонстрував значні переваги у швидкодії порівняно з використанням лише адверсаріального методу.

Особливо важливо підкреслити критичну роль попередньої анонімізації даних перед їх потраплянням до великих мовних моделей (ВММ). Ця попередня обробка є ключовим етапом у забезпеченні належного визначення та захисту персональних даних. Без такої попередньої анонімізації існує значний ризик витоку чутливої інформації через ВММ, що може призвести до серйозних порушень приватності та потенційних юридичних наслідків.

Основні досягнення включають зменшення часу обробки одного коментаря на 40%, підвищення точності анонімізації медичних даних на 5%, покращення точності анонімізації ідентифікуючих даних на 6%, та збільшення якості збереження корисності медичних даних на 7%. Експериментальне дослідження на наборі з 10,000 коментарів підтвердило ефективність двоетапного підходу, де k-анонімність успішно обробила 8,000 елементів персональних даних на першому етапі, а адверсаріальний метод ефективно опрацював решту 4,000 складних випадків.

Запропонований метод відповідає сучасним вимогам захисту персональних даних, забезпечуючи відповідність стандартам GDPR та CCPA, зберігаючи високу якість медичних даних після анонімізації та ефективно протидіє відомим методам деанонімізації та атакам на основі змагальних прикладів.

Однак, були виявлені певні обмеження та напрямки подальшого вдосконалення методу, включаючи необхідність оптимізації обчислювальних ресурсів для обробки великих наборів даних, потребу в додатковому тестуванні на різних типах медичних даних, та можливість адаптації методу для інших галузей застосування.

Практичне значення отриманих результатів полягає у можливості їх безпосереднього впровадження в системи, що використовують великі мовні моделі для обробки медичних даних. Розроблений метод дозволяє значно підвищити рівень захисту персональної інформації при збереженні високої швидкодії та якості даних.

Подальші дослідження можуть бути спрямовані на розробку адаптивних механізмів налаштування параметрів анонімізації, дослідження можливостей масштабування методу для обробки надвеликих наборів даних, розширення застосування методу на інші типи чутливих даних, та вдосконалення механізмів протидії новим типам атак на приватність.

Список літератури

[1] California Consumer Privacy Act (CCPA). [Online]. Available: https://oag.ca.gov/privacy/ccpa. Accessed: Oct. 2018.

[2] EU, “General data protection regulation,” 2016. [Online]. Available: https://gdpr-info.eu. Accessed: Oct. 2024.

[3] U.S. Department of Labor, “DOL,” 2023. [Online]. Available: https://www.dol.gov/general/ppii. Accessed: Oct. 2024.

[4] N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramer, and C. Zhang, “Quantifying memorization across neural language models,” arXiv:2202.07646, Mar. 06, 2023. doi: 10.48550/arXiv.2202.07646.

[5] S. Vimercati, S. Foresti, G. Livraga, and P. Samarati, “k-Anonymity: From Theory to Applications,” Trans. Data Priv., 2023. [Online]. Available: https://www.tdp.cat/issues21/tdp.a460a22.pdf. Accessed: Oct. 23, 2024.

[6] “Differential privacy for deep and federated learning: A survey,” IEEE Access, vol. 10, pp. 8602–8616, 2022. doi: 10.1109/ACCESS.2022.3151670. Accessed: Oct. 16, 2024.

[7] Y. Zhao and J. Chen, “A survey on differential privacy for unstructured data content,” ACM Comput. Surv., vol. 54, no. 10s, pp. 207:1–207:28, Sep. 2022. doi: 10.1145/3490237.

[8] P. R. Silva, J. Vinagre, and J. Gama, “Towards federated learning: An overview of methods and applications,” WIREs Data Min. Knowl. Discov., vol. 13, no. 2, p. e1486, 2023. doi: 10.1002/widm.1486.
[9] J. Li, Y. Yang, Z. Wu, V. G. Vydiswaran, and C. Xiao, “ChatGPT as an attack tool: Stealthy textual backdoor attack via blackbox generative model trigger,” arXiv:2304.14475, 2023. doi: 10.48550/arXiv.2304.14475.
[10] DPIA, 2019. [Online]. Available: https://gdpr.eu/wp-content/uploads/2019/03/dpia-template-v1.pdf. Accessed: Oct. 2024.
[11] R. Staab, M. Vero, M. Balunović, and M. Vechev, “Large language models are advanced anonymizers,” arXiv:2402.13846, 2024. [Online]. Available: https://arxiv.org/abs/2402.13846. doi: 10.48550/arXiv.2402.13846. Accessed: Oct. 03, 2024.
[12] R. Staab, M. Vero, M. Balunović, and M. Vechev, “Beyond memorization: Violating privacy via inference with large language models,” arXiv:2310.07298, May 06, 2024. [Online]. Available: http://arxiv.org/abs/2310.07298. doi: 10.48550/arXiv.2310.07298. Accessed: Oct. 03, 2024.

LARGE LANGUAGE MODELS AND PERSONAL INFORMATION: SECURITY CHALLENGES AND SOLUTIONS THROUGH ANONYMIZATION

Zamroz P.I., Morozov Y.V.

Lviv Polytechnic National University Department of Electronic Computing Machines

E-mail: pavlo.i.zamroz@lpnu.ua, yurii.v.morozov@lpnu.ua

In light of the growing capabilities of Large Language Models (LLMs), there is an urgent need for effective methods to protect personal data in online texts. Existing anonymization methods often prove ineffective against complex LLM analysis algorithms, especially when processing sensitive information such as medical data. This research proposes an innovative approach to anonymization that combines k-anonymity and adversarial methods. Our approach aims to improve the efficiency and speed of anonymization while maintaining a high level of data protection. Experimental results on a dataset of 10,000 comments showed a 40% reduction in processing time (from 250 ms to 150 ms per comment) compared to traditional adversarial methods, a 5% improvement in medical data anonymization accuracy (from 90% to 95%), and a 7% improvement in data utility preservation (from 85% to 92%). Special attention is paid to the application of the method in the context of interaction with LLM-based chatbots and medical information processing. We conduct an experimental evaluation of our method, comparing it with existing industrial anonymizers on real and synthetic datasets. The results demonstrate significant improvements in both data utility preservation and privacy protection. Our method also takes into account GDPR requirements, setting a new standard in the field of data anonymization for AI interactions. This research offers a practical solution for protecting user privacy in the era of LLMs, especially in sensitive areas such as healthcare.

Keywords: AI, data security, ML, LLM, privacy.