Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)
Кафедра Штучного інтелекту
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

Метод автоматизованого визначення параметра ε у диференційній конфіденційності з використанням баєсівскої моделі ризику
(тема)

Виконав:
здобувач другого року навчання,
групи ДСм-24-1
Кобилянська О.А.
(прізвище, ініціали)
Спеціальність 122 Комп'ютерні науки
(код і повна назва спеціальності)
Тип програми освітньо-професійна
(освітньо-професійна або освітньо-наукова)
Освітня програма Науки про дані (Data Science)
(повна назва спеціалізації)
Керівник проф. Філатов В.О.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри _____ Лариса ЧАЛА
(підпис) (прізвище, ініціали)

2025 р.

Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)

Кафедра Штучного інтелекту
(повна назва)

Рівень вищої освіти другий (магістерський)

Спеціальність 122 Комп'ютерні науки
(код і повна назва)

Тип програми освітньо-професійна
(освітньо-професійна або освітньо-наукова)

Освітня програма Науки про дані (Data Science)
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«_____» _____ 20 ___ р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувачеві Кобилянській Олені Андріївні
(прізвище, ім'я, по батькові)

1. Тема роботи Метод автоматизованого визначення параметра ε у диференційній конфіденційності з використанням баєсівської моделі ризику
затверджена наказом університету від 24 листопада 20 25 р. № 1057Ст
2. Термін подання студентом роботи до екзаменаційної комісії 16 грудня 20 25 р.
3. Вихідні дані до роботи Нормативні документи та рекомендації з диференційної приватності й захисту персональних даних; наукові публікації щодо атак на приналежність до вибірки та ризик-орієнтованих моделей; відкриті табличні, візуальні та текстові набори даних для навчання моделей машинного навчання; базові архітектури моделей та їх неприватні конфігурації; результати попередніх експериментів з DP-SGD і MIA, що використовуються як вихідні орієнтири для побудови та калібрування моделі ризику.

4. Перелік питань, що потрібно опрацювати в роботі
1) Стан диференційної конфіденційності та автоматизованого визначення параметра епсілон
2) Моделювання та практична реалізація автоматизованого вибору епсілон
3) Емпіричний аналіз атак на приналежність, механізмів захисту та автоматизованого вибору параметра епсілон

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Отримання завдання на кваліфікаційну роботу	24.11.2025	виконано
2	Аналіз нормативних та наукових джерел	25.11.2025	виконано
3	Розробка теоретичної моделі ризику системи	26.11.2025	виконано
4	Проєктування експериментальної методики та	28.11.2025	виконано
	даних		
5	Реалізація програмних модулів і алгоритмів	29.11.2025	виконано
6	Проведення експериментів і обробка результатів	30.11.2025	виконано
7	Написання та оформлення пояснювальної	01.12.2025	виконано
	записки		
8	Перевірка на академічний плагіат	03.12.2025	виконано
9	Підготовка презентації та доповіді	06.12.2025	виконано
10	Попередній захист	08.12.2025	виконано
11	Рецензування	09.12.2025	виконано
12	Захист перед ЕК	16.12.2025	

Дата видачі завдання 24 листопада 2025 р.

Здобувач _____
(підпис)

Керівник роботи _____ проф. Філатов В.О.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка: 82 с., 8 рис., 8 табл., 1 дод., 21 джерело.

АВТОМАТИЗОВАНИЙ ВИБІР ПАРАМЕТРА ЕПСІЛОН, АТАКИ НА ПРИНАЛЕЖНІСТЬ ДО ВИБІРКИ, БЮДЖЕТ ПРИВАТНОСТІ, ДИФЕРЕНЦІЙНА КОНФІДЕНЦІЙНІСТЬ, ЗАХИСТ ДАНИХ, КОРИСНІСТЬ МОДЕЛІ.

Об’єкт дослідження – процес навчання моделей машинного навчання над чутливими даними за умов диференційної конфіденційності.

Предмет дослідження – методи кількісного оцінювання ризику атак на приналежність до вибірки та корисності моделей, а також підходи до автоматизованого вибору параметра ε для механізмів диференційної конфіденційності.

Мета роботи – розробка та експериментальна перевірка ризику орієнтованого підходу до вибору параметра ε, який узгоджує вимоги приватності та якості моделей машинного навчання на основі байєсівської моделі ризику й емпіричних метрик успішності атак.

Методи дослідження – теоретичний аналіз моделей диференційної конфіденційності та атак на приналежність до вибірки, побудова й дослідження байєсівських імовірнісних моделей, статистична обробка результатів експериментів, моделювання loss- та shadow-атак, аналіз кривих «ризик – корисність» для різних значень ε.

У результаті роботи запропоновано інтегральну модель ризику, яка поєднує внесок різних типів атак у єдиній нормованій шкалі, та композитну функцію втрат, що одночасно враховує емпіричний ризик витоку та втрату корисності. Побудовано процедуру автоматизованого вибору ε як аргументу мінімуму цієї функції з можливістю налаштування пріоритетів між приватністю та якістю моделі.

ABSTRACT

Master’s thesis contains: 82 pp., 8 fig., 8 tabl., 1 ann., 21 references.

AUTOMATED EPSILON SELECTION, DATA PROTECTION, DIFFERENTIAL PRIVACY, MEMBERSHIP INFERENCE ATTACKS, MODEL UTILITY, PRIVACY BUDGET.

Object of the study – the process of training machine-learning models on sensitive data under differential privacy constraints.

Subject of the study – methods for quantitative assessment of the risk of membership inference attacks and model utility, as well as approaches to automated selection of the parameter ε for differential privacy mechanisms.

Aim of the work – to develop and experimentally validate a risk-oriented approach to selecting the parameter ε that reconciles privacy requirements and model quality, based on a Bayesian risk model and empirical metrics of attack success.

Research methods – theoretical analysis of differential privacy models and membership inference attacks, construction and study of Bayesian probabilistic models, statistical processing of experimental results, simulation of loss and shadow attacks, and analysis of risk–utility curves for different ε values.

As a result of the work, an integral risk model is proposed that combines the contributions of different attack types into a single normalized scale, as well as a composite loss function that jointly accounts for empirical leakage risk and utility loss. A procedure for automated selection of ε is constructed as the argument minimizing this function, with tunable priorities between privacy and model quality.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

БП – бюджет приватності;

ГЧ – глобальна чутливість;

ДК – диференційна конфіденційність;

ДП – диференційна приватність;

ЄС – Європейський Союз;

КДК – концентрована диференційна конфіденційність;

ЛЧ – локальна чутливість;

РД – рівні дивергенції;

AUC – Area Under the Curve – площа під ROC-кривою;

BDP – Bayesian Differential Privacy – байєсівська диференційна приватність;

CNN – Convolutional Neural Network – згорткова нейронна мережа;

DP-SGD – Differentially Private Stochastic Gradient Descent – диференційно-приватний стохастичний градієнтний спуск;

EU AI Act – European Union Artificial Intelligence Act – Акт Європейського Союзу про штучний інтелект;

FPR – False Positive Rate – показник хибних позитивів;

GDPR – General Data Protection Regulation – Загальний регламент захисту даних;

IoT – Internet of Things – Інтернет речей;

LLM – Large Language Model – велика мовна модель;

MIA – Membership Inference Attack – атака на приналежність до вибірки;

MLP – Multilayer Perceptron – багатошаровий перцептрон;

MNIST – Modified National Institute of Standards and Technology database – еталонний набір зображень рукописних цифр;

NIST – National Institute of Standards and Technology – Національний інститут стандартів і технологій;

RAPPOR – Randomized Aggregatable Privacy-Preserving Ordinal Response – рандомізований агрегований протокол збереження приватності відповідей;

RDP – Rényi Differential Privacy – диференційна приватність за Реньї;

ReLU – Rectified Linear Unit – випрямлена лінійна активаційна функція;

SSN – Social Security Number – номер соціального страхування;

TPR – True Positive Rate – показник справжніх позитивів;

zCDP – zero-Concentrated Differential Privacy – нульова концентрована диференційна приватність.

ВСТУП

У сучасних умовах стрімкого зростання обсягів персональних даних та повсюдного впровадження систем штучного інтелекту питання формалізованого захисту приватності набуває визначального значення. Диференційна конфіденційність розглядається як одна з найстрогіших і математично обґрунтованих концепцій захисту, що гарантує обмеження впливу будь-якого окремого запису на результат статистичного або машинного аналізу. Попри значний теоретичний прогрес і поширення відповідних підходів у промислових сервісах, залишається невирішеною прикладна проблема практичного вибору параметра ε, який визначає компроміс між рівнем конфіденційності та корисністю отриманих результатів. У реальних системах застосовуються різні значення цього параметра, що часто ґрунтуються на емпіричному досвіді або внутрішніх політиках організацій, а не на формалізованій оцінці ризиків і якості моделей.

Актуальність роботи зумовлена поєднанням тенденцій розвитку. Сучасні регуляторні акти про захист даних вимагають обґрунтованого використання методів анонімізації та демонстрації залишкового ризику для суб’єктів даних. Паралельно зростає кількість практичних атак на моделі машинного навчання, серед яких окреме місце займають атаки на приналежність до вибірки, здатні виявляти факт участі конкретної особи у навчанні моделі. За відсутності прозорих критеріїв вибору ε виникають такі проблеми, як надмірне зашумлення і втрата корисності або недооцінка реальних вразливостей системи.

Метою кваліфікаційної роботи є розробка та експериментальна перевірка підходу до автоматизованого вибору параметра ε для механізмів диференційної конфіденційності у задачах машинного навчання на основі байєсівської моделі ризику та емпіричних метрик успішності атак на приналежність до вибірки. Для досягнення поставленої мети пропонується

поєднати теоретичні гарантії диференційної конфіденційності, результати практичних атак, зокрема loss та shadow, а також показники якості моделей у єдину композитну функцію, яка кількісно описує компроміс між приватністю та корисністю. У роботі аналізуються як контрольовані синтетичні сценарії, так і реалістичні задачі класифікації табличних, візуальних і текстових даних, що дає змогу дослідити поведінку запропонованого підходу за різних рівнів перенавчання і для різних класів моделей.

Отримані результати можуть бути використані у медичних інформаційних системах, фінансовій аналітиці, державних реєстрах, а також у промислових платформах машинного навчання, де моделі тренуються на чутливих даних користувачів. Запропонована методика здатна слугувати інструментом для проєктування політик приватності, документування прийнятого рівня ризику та автоматизованого налаштування гіперпараметрів навчання.

1 СТАН ДИФЕРЕНЦІЙНОЇ КОНФІДЕНЦІЙНОСТІ ТА АВТОМАТИЗОВАНОГО ВИЗНАЧЕННЯ ПАРАМЕТРА ЕПСІЛОН

Актуальність дослідження методів автоматизованого визначення параметра є зумовлена протиріччям між теоретичною досконалістю диференційної конфіденційності (ДК) як математичної концепції та практичними труднощами її реалізації в реальних системах обробки даних. Диференційна конфіденційність гарантує математично обґрунтований рівень захисту персональних даних шляхом контрольованого додавання статистичного шуму, де параметр є виступає ключовим регулятором компромісу між конфіденційністю та утилітарністю результатів аналізу. Проте відсутність універсально прийнятної методології квантитативного визначення оптимального значення є створює суттєвий бар'єр для масштабного впровадження ДК у промислових системах та державних інформаційних реєстрах.

Проблема вибору є набуває додаткової гостроти в контексті сучасного регулювання захисту даних, зокрема вимог Загального регламенту ЄС про захист даних GDPR та інших національних законодавчих актів, які вимагають доведення адекватності застосованих засобів анонімізації. Традиційні підходи, що базуються на емпіричних рекомендаціях типу $\varepsilon = 0.1$ або $\log(2)$, не враховують специфіку предметної області, розподілу даних, вимог до точності аналітичних запитів та ризиків пов'язаних з можливістю атак на відновлення членства, що призводить до ситуації, де практики змушені або надто консервативно витрачати бюджет приватності, різко знижуючи корисність даних, або навпаки недооцінювати ризики, створюючи ілюзію захисту.

Отже, необхідність розробки автоматизованих методів визначення ε, що інтегрують оцінку ризиків, утилітарні обмеження та байєсівські моделі невизначеності, є критичною для подальшого розвитку практичної криптографії та конфіденційного машинного навчання. Особлива

важливість таких методів проявляється в застосуваннях, де дані містять високочутливу інформацію, таку як медичні записи, фінансові транзакції або геолокаційні дані, де помилки в оцінці ризиків можуть мати катастрофічні наслідки для приватності окремих осіб.

1.1 Сучасний стан диференційної конфіденційності у світі

Історичний розвиток диференційної конфіденційності бере свій початок від фундаментальних праць Сінтії Дворк та її колег у Microsoft Research у 2006 році [1], де була вперше формалізована концепція математичних гарантій приватності, незалежних від наявності додаткової інформації про зловмисника. Перші теоретичні розробки фокусувалися на визначенні строгих меж можливостей отримання інформації про окремих осіб з агрегованих статистичних запитів, що призвело до формулювання ε-диференційної конфіденційності як основної дефініції.

Еволюція концепції відбувалася паралельно з розвитком реальних впроваджень у великих технологічних корпораціях. Google першим масштабно впровадив ДК у 2014 році через систему RAPPOR (Randomized Aggregatable Privacy-Preserving Ordinal Response) [2] для збору статистики використання Chrome браузера без ризику розкриття індивідуальних шаблонів користувачів. Система RAPPOR застосовувала локальну диференційну конфіденційність, де шум додавався безпосередньо на пристрої користувача, що усувало необхідність довіряти центральному серверу обробки даних. Згодом Google розширив застосування ДК на сервіси Google Maps для аналізу завантаженості місць, Google Fi для покращення якості зв'язку та Privacy Sandbox для рекламних технологій, де ε коливався від 2 до 8–9 протягом життєвого циклу даних.

Apple прийняв локальну диференційну конфіденційність як фундаментальну архітектурну принципу збору аналітичних даних у iOS та macOS [3]. Реалізація Apple відрізняється використанням Count Mean Sketch

та приватних хеш-матриць з обмеженим бюджетом приватності на рівні $\varepsilon = 4–8$ для збору даних про використання емодзі, введення тексту та перегляду веб-сторінок. Критично важливим елементом підходу Apple є концепція періодичного скидання бюджету приватності, що запобігає акумуляції ризиків при тривалому спостереженні.

Microsoft інтегрував ДК у безліч продуктів, включаючи Windows Telemetry, LinkedIn Advertiser Queries, suggested replies в Office та менеджерські дашборди, використовуючи як централізовані, так і локальні механізми [4]. Дослідницький підрозділ Microsoft Research продовжує активно розвивати теоретичні основи, зокрема в напрямку баєсівської оцінки гарантій приватності. Uber застосовує ДК для аналізу статистичних тенденцій у базі користувачів без розкриття особистої інформації, що критично важливо для геопросторового аналізу та оптимізації маршрутів [5].

Державне регулювання активно адаптує ДК як механізм досягнення анонімізації відповідно до GDPR [6]. Рекомендації NIST щодо управління ризиками приватності та EU AI Act [7] визначають ДК як пріоритетний технічний захід для систем штучного інтелекту, що обробляють персональні дані. Проте жоден з цих регламентів не встановлює конкретних вимог до значення ε, залишаючи остаточне рішення на розсуд адміністраторів даних, що підсилює необхідність науково обґрунтованих методів вибору.

На рисунку 1.1 показано динаміку публікацій у ключових доменах застосування ДК, а саме захист локаційних даних, текстової інформації, зображень, графів, голосових і трафікових потоків, а також універсальні та специфічні реалізації у різних галузях. Яскраво видно експоненційне зростання кількості досліджень у сфері аналізу локацій, що узгоджується з впровадженням ДК у вищезазначених сервісах, а також стійкий приріст у напрямках текстуальної обробки для рекомендаційних систем і безпечних чат-ботів.

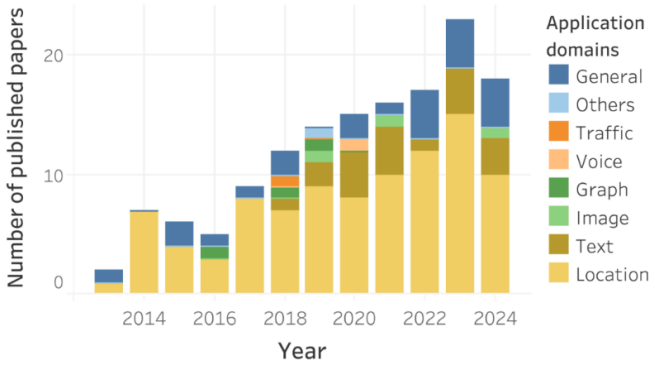

Рисунок 1.1 – Зростання впровадження DP у багатьох сферах застосування [8]

Таблиця 1.1 демонструє порівняльний аналіз значень ε у реалізаціях провідних технологічних компаній, що підкреслює відсутність уніфікованих підходів.

Таблиця 1.1 – Застосування значень ε у технологічних компаніях

Компанія	Застосування	Діапазон ε	Механізм	Рік впровадження
Google	Chrome (malware detection)	2-8.5	Локальний DP	2014
Google	Maps (busyness)	2-4	Централізований DP	2017
Apple	iOS Analytics	4-8	Локальний DP	2016
Microsoft	Windows Telemetry	Не публікується	Гібридний	2015
Uber	User trends analysis	Не публікується	Централізований DP	2017

Не менш важливо зазначити, що починаючи з 2018 року спостерігається розширення сфер застосування ДК на суміжні домени, включаючи графові моделі (аналіз соціальних структур), захист голосових даних (персональні голосові асистенти), а з 2022 року – інтеграцію ДК у обробку трафіку (інтелектуальні транспортні системи), що підкреслю сучасний тренд, який полягає у мультидисциплінарному впровадженні механізмів ДК, від локальної приватності до комплексних захисних систем у штучному інтелекті та IoT.

1.2 Теоретичні основи диференційної конфіденційності та визначення ε

ε-диференційної конфіденційності встановлює суворі математичні гарантії приватності через концепцію інстинктивності сусідніх баз даних. Для механізму M та параметра ε > 0 кажуть, що M задовольняє вимогам ε-ДК якщо для будь-яких двох сусідніх баз даних D і D′, що відрізняються рівно одним записом, та для будь-якої підмножини виводів S ⊆ Range(M) виконується нерівність:

$$Pr[M(D) \in S] \leq e^{\varepsilon} \cdot Pr[M(D') \in S], \quad (1.1)$$

де Pr – ймовірність, що обчислюється по всім випадковим подіям механізму M.

Формула 1.1 формалізує принцип, що впровадження механізму на наборі даних, де присутня одна конкретна особа, не повинно суттєво змінювати розподіл ймовірностей виходів в порівнянні з ситуацією, де цю особу вилучено з набору.

Величина ε називається параметром бюджету приватності та контролює граничне відношення ймовірностей. На логарифмічній шкалі це означає, що максимальна різниця між $log(Pr[M(D) \in S])$ та $log(Pr[M(D') \in S])$ обмежена параметром ε. Математично записується як:

$$| \log(Pr[M(D) \in S]) - \log(Pr[M(D') \in S]) | \leq \varepsilon. \quad (1.2)$$

Менші значення ε забезпечують сильніші гарантії приватності, оскільки розподіли M(D) та M(D′) стають статистично близькими. Однак призводять до необхідності додавання більшого шуму, що знижує точність результатів аналізу.

Варіант ε-диференційної конфіденційності називають чистою ДК. Проте на практиці частіше використовується послаблена версія, звана (ε, δ)-ДК, що допускає малу ймовірність δ порушення гарантій:

$$Pr[M(D) \in S] \leq e^{\varepsilon} \cdot Pr[M(D') \in S] + \delta. \quad (1.3)$$

Параметр δ представляє максимальну ймовірність того, що механізм M не задовольняє суворому визначенню ε-ДК, проте допускають заради значного підвищення утилітарності. На практиці δ зазвичай встановлюється як $\frac{1}{n^2}$, де n – розмір датасету.

Важливо зазначити, що (ε, δ)-ДК визначає два типи сценаріїв:

– з ймовірністю (1 − δ): забезпечуються строгі гарантії ε-ДК;

– з ймовірністю δ: гарантій немає взагалі.

1.2.1 Концепція чутливості функції та калібрування шуму

Чутливість функції f є критичним параметром, що визначає кількість шуму, необхідного для досягнення бажаного рівня ε. Глобальна чутливість (ГЧ) функції f визначається як:

$$\Delta f = max_{D,D\prime:d(D,D\prime)\leq 1} \| f(D) - f(D') \|_1, \quad (1.4)$$

де d(D, D′) ≤ 1 означає, що набори D та D′ відрізняються не більше ніж одним записом;

$\|\cdot\|_1$ позначає L1-норму.

Максимум береться по всім можливим парам сусідніх баз даних, включаючи найгірші сценарії. Для простих функцій глобальна чутливість обчислюється як:

– COUNT має чутливість Δ = 1, оскільки додавання чи вилучення одного запису змінює результат максимум на 1;

– SUM з необмеженими значеннями має чутливість Δ = ∞, оскільки один запис може мати довільно велике значення;

– = COUNT, округлена до найближчого кратного 5, має чутливість Δ = 5.

Але ГЧ часто є надмірно консервативною, оскільки розглядає найгірший можливий сценарій для будь-якої конфігурації даних, включаючи такі, що майже ніколи не трапляються на практиці. Альтернативою є локальна чутливість (ЛЧ):

$$LS_f(D) = max_{D\prime:d(D,D\prime)\leq 1} \mid f(D) - f(D') \mid. \quad (1.5)$$

ЛЧ вимірює максимальну зміну для конкретного набору даних D, а не для всіх можливих наборів. Наприклад, для функції медіани ГЧ може бути нескінченною (якщо немає верхнього обмеження на значення), але на реальних даних ЛЧ часто невелика, оскільки медіана малочутлива до змін даних на краях розподілу. Зв'язок між ними полягає в тому, що для будь-якого набору даних D ЛЧ завжди буде не більша від глобальної.

Використання локальної чутливості вимагає спеціальних механізмів, таких як smooth sensitivity, щоб забезпечити дотримання ε-ДК, оскільки сама величина LS(D) може розкрити інформацію про набір даних.

1.2.2 Механізми Лапласа та Гауса

Найфундаментальнішим механізмом досягнення ε-ДК є механізм Лапласа. Він працює за принципом додавання випадкового шуму, розподіленого згідно розподілу Лапласа:

$$M(D) = f(D) + Y, \quad (1.6)$$

де $Y \sim Lap(\frac{\Delta f}{\varepsilon})$ з масштабним параметром $b = \frac{\Delta f}{\varepsilon}$.

Розподіл Лапласа з параметром масштабу b характеризується щільністю ймовірності:

$$p(y) = \frac{1}{2b} exp\left(-\frac{|y|}{b}\right). \quad (1.7)$$

Математична гарантія полягає в тому, що для будь-яких сусідніх наборів D та D′:

$$\frac{Pr[M(D)=z]}{Pr[M(D\prime)=z]} = \frac{p(z-f(D\prime))}{p(z-f(D))} = exp(-\frac{|z-f(D)|-|z-f(D\prime)|}{b}) \leq exp(\frac{\Delta f}{b}) = e^{\varepsilon}. \quad (1.8)$$

Головним спостереженням є те, що дисперсія доданого шуму пропорційна квадрату масштабу: $Var[Y] = 2b^2 = 2(\frac{\Delta f}{\varepsilon})^2$. Отже, помилка механізму зростає як O($(\frac{\Delta f}{\varepsilon})^2$), а означає це те, що зменшення ε вдвічі призводить до четирикратного зростання помилки.

Для багатовимірних запитів або коли можна допустити малу ймовірність порушення гарантій, часто використовується гауссівський механізм. Він додає шум, розподілений згідно нормального розподілу з дисперсією, масштабованою до L2-чутливості:

$$M(D) = f(D) + Y, Y \sim N(0, \sigma^2 I_d), \quad (1.9)$$

де L2-чутливість визначається як:

$$\Delta_2 f = max_{D,D\prime:d(D,D\prime)\leq 1} \| f(D) - f(D') \|_2. \quad (1.10)$$

Гауссівський механізм задовольняє (ε, δ)-диференційну конфіденційність, і його похибка (в термінах L2-норми) становить O($(\frac{\Delta_2 f}{\varepsilon})^2$),

що подібно до механізму Лапласа. Однак гауссівський шум менше вироджується у хвостах розподілу, що робить його практичніше для багатовимірних аналізів та глибоких нейронних мереж.

1.2.3 Композиція ε та накопичення бюджету приватності

Одна з найважливіших властивостей диференційної конфіденційності – це можливість композиції, яка дозволяє обчислювати сукупне витрачання бюджету приватності при множинних запитах до того ж датасету. Послідовна композиція визначається наступним чином. Якщо механізм M_1 задовольняє ε_1-ДК, а механізм M_2 задовольняє ε_2-ДК, то комбінований механізм $G = (M_1, M_2)$, що застосовує обидва послідовно до того ж датасету, задовольняє $(\varepsilon_1 + \varepsilon_2)$-ДК:

$$Pr[G(D) \in S] \leq e^{\varepsilon_1+\varepsilon_2} \cdot Pr[G(D') \in S]. \quad (1.11)$$

При k запитах, кожен з яких задовольняє ε_0-ДК, сукупний бюджет становить $k \cdot \varepsilon_0$. Тому, якщо хочемо обмежити сукупний бюджет на рівні ε при k запитах, кожен запит отримує бюджет $\varepsilon_0 = \frac{\varepsilon}{k}$.

При паралельній композиції датасет розділено на k непересічних підмножин, і кожна підмножина запитується з бюджетом ε, то сукупний бюджет все ще становить ε (не $k \cdot \varepsilon$), тому що кожен запис у датасеті впливає лише на один з k запитів.

Проте просте додавання ε при послідовній композиції часто є надмірно консервативним. Більш точні результати дають розширені теорії композиції, які використовують рівні дивергенції (РД) або концентровану диференційну конфіденційність (КДК).

КДК вводить нові параметри приватності на основі логарифмічного моменту розподілу втрат приватності, що дозволяє розглядати приватність не як суворе гарантоване обмеження на ймовірності для кожного запиту, а

як статистичне середнє по всьому набору можливих запитів. Основна ідея КДК – замість параметра ε для всіх запитів, враховувати середні та дисперсійні значення. За рахунок чого суттєво покращується адитивність при композиції, і дає менш консервативні оцінки, тому підвищується корисність даних [9].

Порівнюючи механізми ДК, представленими у таблиці 1.2, ключовими критеріями виступають тип чутливості, шуму, порядок помилки, форма та практична оптимальність для конкретних задач. Основна переваг Лапласівського механізму – математична простота, але для складних багатовимірних запитів він має обмеження щодо точності. У такому випадку краще пристосований Гаусівський механізм завдяки L2 чутливості [10]. Геометричний механізм призначений для дискретних запитів, наприклад у підрахунку частот. Змішаний тип поєднує переваги двох підходів та балансує між чутливостями, проте поступається простотою математичного аналізу [11].

Таблиця 1.2 – Характеристики основних механізмів досягнення ДК

Механізм	Чутливість	Тип шуму	Порядок помилки	Форма	Використання
Лапласа	L1 (Δ)	Лапласа	O(Δ/ε)	ε-ДК	Оптимальний для одновимірних запитів
Гауссівський	L2 (Δ_2)	Гауссів	O(Δ_2/ε)	(ε,δ)-ДК	Краще для багатовимірних
Геометричний	L0	Геометричний	O(1/ε)	ε-ДК	Для дискретних запитів
Гібридний Huber	L1/L2 гібрид	Huber	O(Δ/ε)	ε-ДК	Компроміс між Лапласом та Гаусом

Сучасні механізми РД і КДК суттєво полегшують роботу з великою кількістю композиційних запитів, забезпечуючи більш гнучке управління

накопиченим бюджетом приватності, що є критичним для довготривалих аналітичних процесів і інтеграції у складні моделі глибокого навчання. Такі механізми дають практику фактично нарощувати масштаб та зберігати високий рівень захисту, контролюючи лише статистичне середнє втрат приватності замість кожної дії окремо.

1.3 Підходи до вибору ε, компроміси та недоліки

Парадокс компромісу в диференційній конфіденційності полягає в неможливості одночасно досягти максимальної утилітарності даних та абсолютної конфіденційності окремих записів. Кожен запит до ДК-системи витрачає бюджет приватності (БП), поступово знижуючи гарантії захисту для наступних операцій. На основі цього створюється проблематика для практиків, які мають визначити прийнятну кількість шуму для конкретного аналітичного завдання. Основні виклики інтерпретації ε включають:

– відсутність універсальної шкали оцінки ризиків, де ε = 1 не може бути однозначно кваліфіковано як «безпечно» або «небезпечно» без контексту даних;

– накопичення помилок при композиції запитів, де сукупний ризик зростає лінійно, або сублінійно при використанні розширених методів обліку, з кількістю операцій;

– несприйнятливість кінцевих користувачів до абстрактних математичних гарантій, що вимагає перекладу ε в понятні категорії ризиків;

– необхідність балансування між інтересами аналітика даних, який прагне мінімізувати шум, та учасника дослідження, який вимагає максимального захисту.

Практичні дослідження показують, що навіть досвідчені фахівці з ДК мають складнощі з обґрунтованим вибором ε. Дослідження практиків, що впроваджують ДК, виявило широкий діапазон використовуваних значень, від ε = 0.01 для високочутливих медичних даних до ε = 15–17 для

демографічних статистик перепису населення. Таке розмаїття свідчить про відсутність стандартизованих методологій та необхідність розроблення адаптивних підходів.

Тому подальші дослідження фокусуються на емпіричній оцінці таких взаємозв'язків. Традиційні методи вибору ε можна класифікувати на декілька категорій, кожна з яких має суттєві обмеження. Емпіричні рекомендації та правила великого пальця базуються на накопиченому досвіді дослідницької спільноти. Ранні поради рекомендували ε в діапазоні [0.1; 1] для більшості застосувань [12], тоді як деякі джерела пропонували log(2) або log(3) для завдань з помірною чутливістю[13]. Проте ці рекомендації є суто емпіричними, не враховують специфіки даних, вартості помилок або юридичних вимог. Недоліком є їхня невідповідність принципу ризик-орієнтованого підходу, що вимагається GDPR [6].

Економічні моделі пропонують інтерпретувати ε через призму вартості приватності та компенсацій учасникам. Модель Хсу [13] та співавторів використовує параметри, такі як вартість даних для аналітика, ризики для учасника та вартість компенсації, для розрахунку оптимального ε, що максимізує соціальну корисність.

Методи, засновані на вимогах до точності, визначають ε залежно від допустимого рівня похибки в аналітичних запитах. Для заданого функціоналу чутливості Δf та допустимої варіації результатів можна обчислити необхідний масштаб шуму, що однозначно визначає ε. Проте ігноруються ризики приватності, фокусуючись виключно на утилітарності, що може призвести до недооцінки загроз.

Методи, орієнтовані на регуляторні вимоги, використовують ε як засіб досягнення статусу анонімізованих даних відповідно до GDPR. Деякі автори аргументують, що ДК з достатньо малим ε може задовольняти вимоги статті 26 GDPR щодо анонімізації [14]. Проте регулятори навмисно не встановлюють конкретних порогових значень ε, оскільки прийнятність залежить від контексту, що створює юридичну невизначеність.

Усі методи мають спільний недолік: вони не враховують динамічної природи ризиків та не забезпечують механізмів адаптації є до змін у розподілі даних, архітектурі моделей або векторах атак.

1.4 Membership Inference Attack як міст між теорією і практикою

Membership Inference Attack (MIA) є інструментом для оцінки емпіричної ефективності гарантій диференційної конфіденційності. Атака на відновлення членства передбачає, що зловмисник, маючи доступ до «чорного ящика» моделі чи її виходів, намагається визначити, чи був конкретний запис даних використаний під час навчання цільової моделі. Успішність такої атаки безпосередньо вимірює, наскільки модель «запам'ятовує» індивідуальні зразки, що суперечить принципам ДК.

Теоретичний зв'язок між MIA та ДК базується на спостереженні, що механізми з сильними гарантіями повинні знижувати дискримінативну здатність атакуючої моделі до рівня вгадування. Для аналізу використовують метрики, такі як AUC або advantage. У ідеальному випадку ДК-захищена модель повинна зменшити advantage до умовного нуля, хоча на практиці досягається це ціною значного зниження точності моделі [15].

MIA як метод оцінки має суттєві обмеження. По-перше, ефективність атаки сильно залежить від архітектури цільової моделі, розміру навчальної вибірки та розподілу даних. Глибші нейронні мережі з великою кількістю параметрів схильні до більшого перенавчання, що знижує їхню вразливість до MIA незалежно від значення ε. По-друге, традиційні MIA припускають пасивного зловмисника, тоді як активні атаки, що підтримують навчання зловмисної моделі, можуть досягати значно вищої ефективності.

Критичним недоліком є те, що MIA оцінює приватність конкретної реалізації моделі, а не абстрактного механізму. Виходить, що атака може показати високу вразливість системи, яка задовольняє вимогам ε-ДК, через неоптимальну реалізацію або недостатній облік допоміжної інформації.

Хоча модель зі слабким є може виявитися стійкою до конкретних атак через специфіку даних.

Незважаючи на ці обмеження, MIA залишається незамінним інструментом для практичної оцінки, оскільки вона перетворює абстрактні математичні гарантії в конкретні метрики ризику. Сучасні підходи використовують MIA як компонент внутрішньої оцінки приватності, де результати атаки інтегруються в байєсівські моделі для отримання постеріорних розподілів реальних ризиків.

На рисунку 1.2 представлена послідовність дій, які виконує зловмисник для визначення, чи певний запис належить до навчальної вибірки цільової моделі f_target. Зловмисник генерує запит – подає певний зразок даних на вхід цільової моделі та отримує її прогноз. Далі цей прогноз разом із реальним класом чи міткою використовується для навчання або використання моделі атаки f_attack.

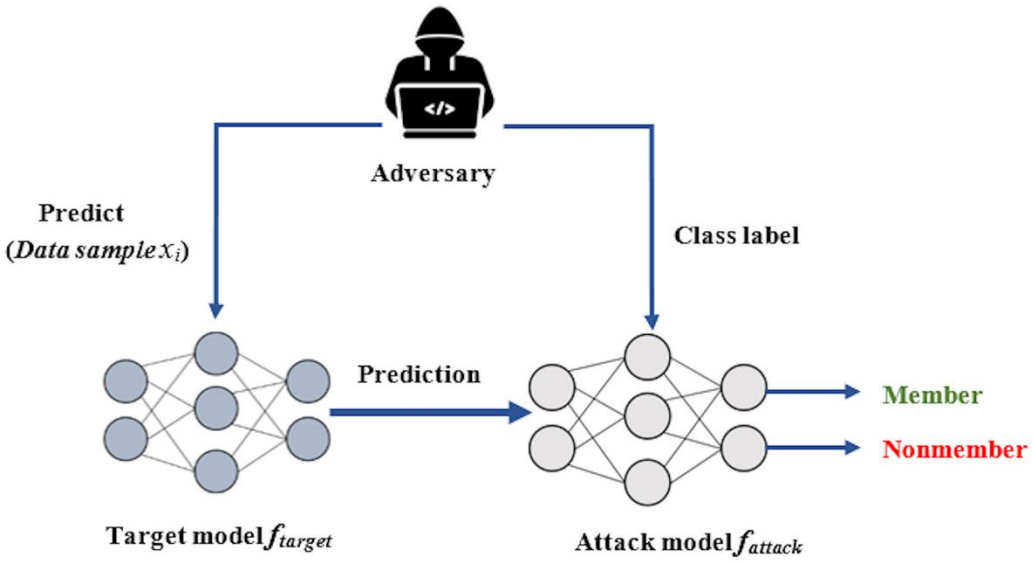

Рисунок 1.2 – Membership Inference Attack [15]

Атакуюча модель аналізує вихід цільової моделі, а також ознаки зразка, і класифікує результат як «Member» або «Nonmember». Таким чином, головна мета зловмисника – підвищити точність імовірнісної оцінки

членства запису у тренувальному наборі, використовуючи різницю поведінки цільової моделі щодо своїх і сторонніх прикладів.

Не менш важливо, що структура взаємодії дає змогу протестувати рівень приватності конкретної реалізації ДК-механізму: чим менша ймовірність правильного визначення членства, тим сильніші гарантії приватності.

1.5 Байєсівські підходи у задачах приватності та ризику

Байєсівські методи в ДК пропонують комплексний фреймворк для кількісної оцінки невизначеності параметрів ризику та інтеграції апріорних знань про розподіл даних. На відміну від класичного частотного підходу, який трактує ε як фіксовану константу, байєсівська парадигма розглядає його як випадкову величину з певним апріорним розподілом, що оновлюється на основі спостережень за результатами атак або порушень приватності.

Ключовою інновацією є Bayesian Estimation of Differential Privacy, де оцінка гарантій приватності формулюється як задача статистичного висновку. Метод використовує результати MIA для побудови спільного постеріорного розподілу для false positive та false negative rates, що дозволяє отримати повний постеріорний розподіл замість простого довірчого інтервалу. У результаті скорочується необхідний обсяг вибірки для досягнення статистичної значущості на 40% порівняно з частотними методами [16].

Bayesian Differential Privacy (BDP) пропонує ще радикальніший підхід, де гарантії приватності калібруються з урахуванням розподілу даних. Традиційна ДК припускає, що будь-які два сусідні набори даних є рівно можливими, тоді як BDP використовує апріорні знання про ймовірність появи конкретних записів, що дає змогу отримувати значно

тісніші гарантії приватності для типових in-distribution зразків, зберігаючи строгі захисні властивості для аномальних даних [16].

Переваги байєсівських методів проявляються в декількох аспектах. По-перше, вони дозволяють інтегрувати витончені моделі загроз, включаючи обмежених за ресурсами зловмисників або зловмисників з частковим доступом до допоміжної інформації. По-друге, байєсівський фреймворк природно підтримує адаптивне оновлення БП в режимі реального часу, реагуючи на зміни в структурі даних або векторах атак. По-третє, він забезпечує засоби для колективного прийняття рішень, де декілька зацікавлених сторін можуть формулювати свої переваги та досягати консенсусу щодо оптимального рівня ризику.

Критичним викликом є вибір апріорних розподілів, що може суттєво впливати на результати. Неінформативні апріори є консервативними, але можуть призвести до надмірного витрачання БП. Інформативні апріори, засновані на історичних даних або експертних оцінках, більш ефективні, але вимагають ретельної валідації для уникнення впливу суб'єктивних упереджень.

1.5.1 Функції втрат

Функції втрат в контексті ДК служать для визначення компромісу між точністю аналітичних результатів та рівнем захисту приватності. Традиційний підхід розглядає обидві метрики як окремі оптимізаційні задачі, тоді як сучасні методи намагаються інтегрувати їх у єдиний цільовий функціонал, що представлено в таблиці 1.3 [17].

Utility loss зазвичай вимірюється як L2-відстань між істинним результатом запиту та ДК-захищеним виводом:

$$L_{utility} = \left|\left|f(D) - M(D)\right|\right|^2. \quad (1.12)$$

Privacy loss може бути записано через ймовірність успішної атаки на відновлення членства:

$$L_{privacy} = \frac{\log(Adv(M)}{Adv_{random}}, \quad (1.13)$$

де Adv(M) – advantage атаки проти механізму M.

Сучасні підходи пропонують комбіновані функції втрат, такі як

$$L_{total} = \alpha \cdot L_{utility} + (1 - \alpha) \cdot L_{privacy}, \quad (1.14)$$

де α ∈ є коефіцієнтом, що відображає пріоритети системи.

Таблиця 1.3 – Порівняльна таблиця функції витрат

Тип функції втрат	Формула	Переваги	Недоліки	Застосування
L2 Utility Loss	||f(D)-M(D)||²	Простота, інтерпретованість	Ігнорує ризики	Агрегатні запити
Logistic Privacy Loss	log(1+Adv)	Враховує реальні атаки	Складність оцінки Adv	Модельні оцінки
Combined Weighted Loss	α·L_utility + β·L_privacy	Гнучкість пріоритетів	Проблема вибору α, β	Адаптивні системи
Pareto Frontier	Multi-objective	Повна картина компромісу	Висока обчислювальна складність	Стратегічне планування

Подібні функції є основою для автоматизованого вибору ε у емпіричних експериментах. Більш тонкі підходи використовують multi-objective optimization, де будується повний фронт Парето рішень, що демонструють всі можливі баланси між утилітарністю та приватністю.

1.6 Байєсівська модель ризику

Визначення успішності атак як випадкової змінної ризику починається з розуміння того, що реальна успішність атак на членство, атрибути чи повторну ідентифікацію залежить від багатьох факторів, які важко передбачити заздалегідь, а саме специфіка розподілу даних, архітектура моделі, навички атакуючого, обсяг допоміжної інформації. Замість того щоб трактувати успішність атак R(ε) як детерміновану функцію ε, байєсівський підхід розглядає R(ε) як випадкову величину з розподілом ймовірностей P(R | ε). Його можна оновити, проводячи експерименти з атаками та спостерігаючи реальні результати.

Апріорний розподіл P(R) відображає попереднє переконання експерта про ймовірний рівень ризику для певного параметра ε до того, як будь-які емпіричні експерименти проведені. На практиці апріор можна обрати на основі теоретичних верхніх меж, досвіду з подібних систем, чи регуляторних рекомендацій. Наприклад, якщо розраховується, що успішність атаки буде дуже низькою при ε = 0.1, то апріорний розподіл повинен бути зосереджений біля низьких значень ризику.

Коли проводяться експерименти атак, результати спостерігаються як послідовність успіхів і невдач атак. Дано N експериментів, під час яких k із них завершилися успіхом, тобто атака вдало виявила членство, передбачила атрибут чи переідентифікувала індивід. Задача байєсівського оновлення полягає у перетворенні апріорного розподілу P(R) у апостеріорний розподіл P(R | D_atack) через теорему Байеса:

$$P(R \mid D_{attack}) = \frac{P(D_{attack}) \cdot P(R)}{P(D_{attack} \mid R)}, \tag{1.15}$$

де D_{attack} – це дані експериментів атак (послідовність успіхів та невдач);

$P(R \mid D_{attack})$ – правдоподібність спостереженого результату при гіпотетичному рівні ризику R;

знаменник $P(D_{attack} \mid R)$ – нормалізуюча константа.

Правдоподібність $P(D_{attack} \mid R)$ моделюється як біноміальне розподілення. Якщо успішність однієї атаки підпорядковується розподілу Бернуллі з параметром R (вірогідність успіху), то при N незалежних випробуваннях вероятність спостерігати рівно k успіхів дорівнює:

$$P(D_{attack} \mid R) = \binom{N}{k} R^k (1-R)^{N-k}. \quad (1.16)$$

Розподіл $P(D_{attack} \mid R)$ отримується множенням апріори на правдоподібність та нормалізацією. Для практичних розрахунків часто вибирається такий розподіл, який є спряженим до біноміального, що дозволяє отримати закриту форму апостеріорного розподілу . Beta-розподіл є таким спряженим для біноміального розподілу, що значно спрощує обчислення. На рисунку 1.3 зображено процес апостеріорного оновлення як еволюцію розподілу ймовірностей при збільшенні кількості спостережень з експериментів атак.

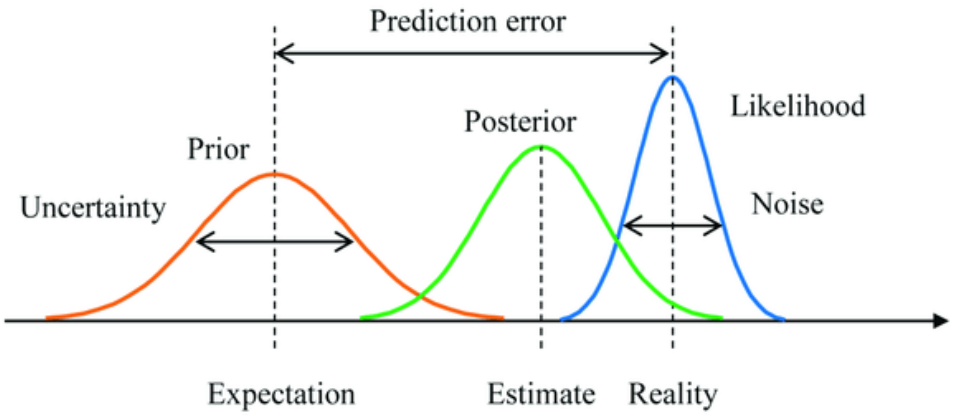

Рисунок 1.3 – Процес апостеріорного оновлення [18]

Не менш важливо зазначити, що вибір апріорного розподілу істотно впливає на апостеріорні висновки, що представлено в таблиці 1.4, особливо при малій кількості експериментів. Неінформативний апріор Beta(1, 1) = Uniform припускає повну невизначеність, тоді як інформативні апріори на кшталт Beta(2, 5) можуть зміщувати оцінку в бік низького ризику. Саме тому процес вибору апріора повинен бути прозорим та документованим, з розглядом альтернативних апріор для перевірки чутливості результатів.

Таблиця 1.4 – Зміна R залежно від кількості невдач та успіхів атак

Апріор	N спроб	k успіхів	E[R | D]	Var[R | D]
Beta(1,1)	10	3	0.375	0.020
Beta(1,1)	10	7	0.727	0.019
Beta(1,1)	100	30	0.300	0.002
Beta(2,5)	10	3	0.348	0.017
Beta(2,5)	100	30	0.297	0.002

Отже, узагальнюючи все вищенаведене, можна стверджувати, що за останні двадцять років ДК пройшла шлях від суто теоретичної конструкції у межах криптографії до практичного інструмента, вбудованого у промислові рішення провідних технологічних компаній. Масштабні проєкти Google, Apple, Microsoft, Uber та інших гравців наочно продемонстрували, що механізми ДК здатні стабільно працювати у високонавантажених системах, які щоденно обробляють мільйони запитів до чутливих наборів даних. Не менш важливо зазначити, що ключовим бар’єром для ще ширшого впровадження ДК у корпоративній та державній практиці залишається проблема узгодженого вибору параметра ε. Наявні підходи, зокрема орієнтація на емпіричні поради, економічні моделі або методики, прив’язані лише до показників точності, виявляють істотні обмеження щодо адаптивності, теоретичного обґрунтування та можливості масштабування.

Підсумовуючи, можна сказати, що ε-диференційна конфіденційність у поєднанні з механізмами Лапласа і Гаусса та властивостями композиції утворює завершений математичний апарат. Водночас саме на стику цієї формальної досконалості та реальних потреб експлуатації виникає головний розрив. Гарантії ДК задаються для найгіршого можливого сценарію й максимально сильного супротивника, тоді як у реальних умовах можливості атакувальної сторони часто обмежені технічно, організаційно або інформаційно. MIA відіграє роль практичного індикатора, що поєднує абстрактні гарантії ДК з емпіричними вимірюваннями ризику та дозволяє оцінити, наскільки механізм з параметром ε фактично протидіє конкретним сценаріям атак. Водночас обмеження, зокрема припущення про пасивного зловмисника та залежність результатів від архітектури моделі й розподілу даних, демонструють потребу у більш гнучких підходах до оцінювання ризиків, які враховують як структуру даних, так і різні вектори атак.

Виходить, що баєсівські підходи до управління ε формують якісно новий рівень розуміння приватності. На відміну від традиційного підходу, де ε розглядають як фіксовану константу, баєсівський фреймворк дозволяє трактувати цей параметр як об’єкт статистичного висновку з явними апріорними уявленнями про прийнятний ризик та подальшим апостеріорним оновленням. Функції втрат, у яких безпосередньо закладено компроміс між корисністю результатів та ступенем конфіденційності, дають математичний інструментарій для раціонального вибору оптимального рівня ε для кожної конкретної задачі. Таким чином, поєднання баєсівського висновку, емпіричної оцінки ризиків за допомогою атак на відновлення членства та явного моделювання компромісу між утилітарністю й приватністю створює теоретичну основу для переходу від пасивного прийняття ε до активного керування ним як стратегічним параметром системи.

2 МОДЕЛЮВАННЯ ТА ПРАКТИЧНА РЕАЛІЗАЦІЯ АВТОМАТИЗОВАНОГО ВИБОРУ ЕПСІЛОН

Теоретичний апарат диференційної конфіденційності, представлений у першому розділі, створює основу для розробки автоматизованих методів вибору параметра ε, що враховують емпіричні компроміси між приватністю та утилітарністю. Метою другого розділу є практична верифікація теоретичних положень через серію контрольованих експериментів, спрямованих на кількісну оцінку впливу ε на стійкість моделей до атак на відновлення членства та витік конфіденційної інформації.

Дослідження охоплює три типи сценаріїв: синтетичні табличні дані, реальні медичні записи (Texas-100), візуальні дані (CIFAR-10) та великі мовні моделі (GPT-2).

Експериментальна методологія є уніфікованою для всіх розглянутих сценаріїв і передбачає послідовне навчання базової моделі в режимі з навмисно індукованим перенавчанням, подальше оцінювання її вразливості за допомогою loss- та shadow-атак на належність, після чого виконується диференційно-приватне навчання з використанням DP-SGD для низки значень ε у діапазоні [0.3; 25000] та повторне вимірювання показників якості і ризику. Для випадку мовної моделі додатково розглядається аналіз прямого витоку секретних рядків (email-адрес, телефонів, SSN), що більш наближено моделює практичні ризики меморизації тренувальних даних. Отримані емпіричні залежності між параметром ε та парою показників (accuracy, AUC) надалі слугуватимуть основою для поглибленого аналізу у третьому розділі.

2.1 Моделювання атак на синтетичному датасеті

Як було показано у теоретичній частині, формальне значення параметра ε не дозволяє інтуїтивно оцінити фактичний ризик витоку

інформації через модель, тому наступним кроком стало моделювання атак на контрольованому синтетичному датасеті. Для всіх експериментів із синтетичними даними використовувалася одна й та сама базова архітектура – трьохшарова повнозв'язна нейронна мережа (MLP) для задачі бінарної класифікації, фрагмент структури представлено в лістингу 2.1.

Лістинг 2.1 – Архітектура безлайн моделі

```
class MLP(nn.Module):
    def __init__(self, input_dim=100, hidden_dims=[128, 64], num_classes=2):
        super().__init__()
        self.fc1 = nn.Linear(input_dim, hidden_dims[0])
        self.ln1 = nn.LayerNorm(hidden_dims[0])
        self.fc2 = nn.Linear(hidden_dims[0], hidden_dims[1])
        self.ln2 = nn.LayerNorm(hidden_dims[1])
        self.fc3 = nn.Linear(hidden_dims[1], num_classes)
        self.relu = nn.ReLU()
        self.dropout = nn.Dropout(0.2)
```

Отже, вхідний шар має розмірність 100 ознак, далі йдуть два приховані шари на 128 і 64 нейрони відповідно з активацією ReLU і Dropout-регуляризацією, а вихідний шар містить два класи для бінарної класифікації. Використання LayerNorm замість BatchNorm зробило архітектуру сумісною з DP-SGD, оскільки обробка здійснюється на рівні окремих прикладів, а не всередині батча, а також дозволило уникнути нестабільної поведінки статистик батча при додаванні шуму.

На першому етапі було згенеровано штучний датасет за допомогою функції make_classification зі 100-вимірними ознаками, 20 інформативними, 10 надлишковими та 5 повторюваними ознаками, двома класами та декількома кластерами в межах кожного класу. За результатами поділу даних було отримано наступні результати:

– train: 1000 прикладів – для посилення перенавчання;
– validation: 2100 прикладів;
– test: 3000 прикладів.

У розширеному сценарії для тіньових моделей було згенеровано 15000 прикладів з аналогічними параметрами, а дані поділено на чотири частини:

– target train: 2000 зразків (навчання цільової моделі);
– target val: 3250 зразків;
– test set: 4500 зразків (публічний тест);
– shadow pool: 5250 зразків (лише для атакувальника);

Отже, для shadow-атаки цільова модель навчається лише на частині даних, а значний пул тіньових прикладів використовується виключно для побудови атакуючих моделей.

Навчання як звичайної, так і диференційно-приватної моделей реалізовано у єдиній функції train_model, де наявність параметра епсілон визначає, чи буде активовано DP-SGD.

```
def train_model(model, train_loader, val_loader,
    epochs=50, lr=0.001,
    epsilon=None, delta=1e-5,
```

У режимі без диференційної конфіденційності (epsilon=None) модель навчається з використанням стандартного оптимізатора Adam, який реалізує стохастичний градієнтний спуск з адаптивною швидкістю навчання для кожного параметра. У випадку DP-SGD обчислюється noise_multiplier для заданих ε, δ, sample_rate і кількості епох, після чого PrivacyEngine автоматично додає шум до градієнтів та виконує кліпінг за нормою max_grad_norm, що дає можливість у рамках одного коду порівняти класичне навчання й навчання з суворими гарантіями диференційної конфіденційності.

Для сценарію з loss-атакою базова модель навчалася 150 епох із батчем 16 і підвищеною швидкістю навчання з параметром lr=0,01, щоб

цілеспрямовано посилити перенавчання. У розширеному випадку модель навчалась протягом 50 епох з батчем 32. Отримані результати представлені в таблиці 2.1.

Таблиця 2.1 – Навчання моделей на синтетичних даних

	Baseline Loss	Baseline Shadow
Train Acc	1.0000	0.9995
Test Acc	0.8100	0.8289
Gap	0.1900	0.1706

Отже, в обох розглянутих режимах модель демонструє практично максимальну точність на тренувальній вибірці, тоді як на тестових даних фіксується суттєвий розрив між якістю навчання та узагальненням, при цьому значення gap перебуває в інтервалі приблизно 0.17–0.19, що вказує на виражене перенавчання.

2.1.1 Реалізація Loss-атаки

У реалізації досліджується проста loss-атака, яка використовує лише значення функції втрат як ознаку членства. Відповідна функція perform_mia (лістинг 2.2) обчислює вектор індивідуальних втрат для тренувальних та тестувальних записів та далі оцінює AUC для задачі.

Лістинг 2.2 – Функція perform_mia

```
def perform_mia(model, X_train, y_train, X_test, y_test):
    model.eval()
    criterion = nn.CrossEntropyLoss(reduction='none')
    with torch.no_grad():
        train_logits    = model(X_train)
        member_losses       =       criterion(train_logits,
y_train).cpu().numpy()
        test_logits     = model(X_test)
```

Продовження лістингу 2.2

```
            non_member_losses     =     criterion(test_logits,
y_test).cpu().numpy()
        return member_losses, non_member_losses, auc
```

Для базової моделі показники атаки належності мають такий вигляд: середня втрата для членів тренувального набору становить 0,0002 ± 0,0010, тоді як для не-членів – 1,1755 ± 2,9511, що дає різницю середніх на рівні 1,1754. Показник AUC = 0,7309 інтерпретується як здатність атакувальника, який має доступ лише до значень функції втрат, з високою ймовірністю коректно визначати належність конкретного запису до тренувальної вибірки.

2.1.2 Реалізація Shadow-атаки

Для більш реалістичної моделі атак було реалізовано shadow-атаку, яка вважається «золотим стандартом» атак членства. На відміну від попереднього підходу, атакувальник явно імітує поведінку цільової моделі, тренуючи ряд тіньових моделей на власному пулі даних, а потім навчає окремий класифікатор над векторами ймовірностей.

Атакувальний сценарій, як було згадано раніше, побудовано на основі shadow pool, який вважається даними, повністю недоступними цільовій моделі, але відомими атакувальнику. Після проходження по всіх п’яти тіньових моделях отриманий атакуючий датасет містить 26250 записів, де кожен приклад – це вектор ймовірностей (2-вимірний softmax-вихід) та мітка 1 (member) або 0 (non-member). Для побудови атакуючого класифікатора використовувався RandomForestClassifier, який якісно працює з невисоковимірними, але складними нелінійними залежностями в просторах ймовірностей, що представлено у лістингу 2.3.

Лістинг 2.3 – Атакуючий класифікатор

```
attack_model = RandomForestClassifier(
    n_estimators=200,
    max_depth=None,
    random_state=SEED,
    n_jobs=-1,
)
attack_model.fit(X_attack_train, y_attack_train)
```

Застосування цієї моделі до неприватної архітектури показало, що значення AUC для атаки на належність становить близько 0,68, тобто суттєво перевищує рівень випадкового вгадування 0,5, що свідчить про збереження високого ризику витоку інформації навіть у разі використання більш розумного атакувальника, який працює не лише з втратами.

2.1.3 Застосування DP-SGD на синтетичних даних

У продовження аналізу захисту синтетичного датасету було застосовано механізм диференційної конфіденційності, який реалізовано через модифікацію процедури навчання за допомогою DP-SGD. Такий підхід не змінює загальну архітектуру MLP, але обмежує внесок окремих спостережень у процес оновлення ваг і тим самим зменшує ризик витоку інформації про конкретні приклади навчальної вибірки. Механізм захисту ґрунтується на наступних умовах:

– попереднє обрізання градієнтів за L_2-нормою, що забезпечує жорстке обмеження чутливості моделі до будь-якого одного прикладу;

– додавання гаусового шуму до вже обрізаних градієнтів, інтенсивність якого контролюється множником шуму та визначає фактичну «силу» приватності;

– урахування багаторазового застосування цього механізму через теореми композиції, що дозволяє підсумовувати витрати приватності й інтерпретувати їх у термінах загального бюджету (ε, δ).

Ефективність такого механізму захисту оцінювалася через його вплив на стійкість до атак на належність. Для неприватної моделі спостерігалося запам’ятовування навчальних прикладів, коли розподіл значень функції втрат для членів тренувальної вибірки суттєво відрізнявся від розподілу для не учасників, а атака на основі MIA досягала AUC близько 0,73, що вказує на високий ризик витоку.

Після переходу до навчання з DP-SGD доданий шум до градієнтів перешкоджає прямому запам’ятовуванню окремих записів, розподіли втрат стають значно ближчими, а AUC для MIA знижується до інтервалу приблизно 0,59–0,60, тобто лише незначно перевищує рівень випадкового вгадування. Усе це досягається за рахунок помірного погіршення точності класифікації: базова модель забезпечує близько 79,7 % правильних рішень при високій уразливості до MIA, тоді як диференційно-приватна конфігурація з ε≈1,0 демонструє близько 74,0 % точності за суттєво зменшеного ризику витоку (рисунок 2.1).

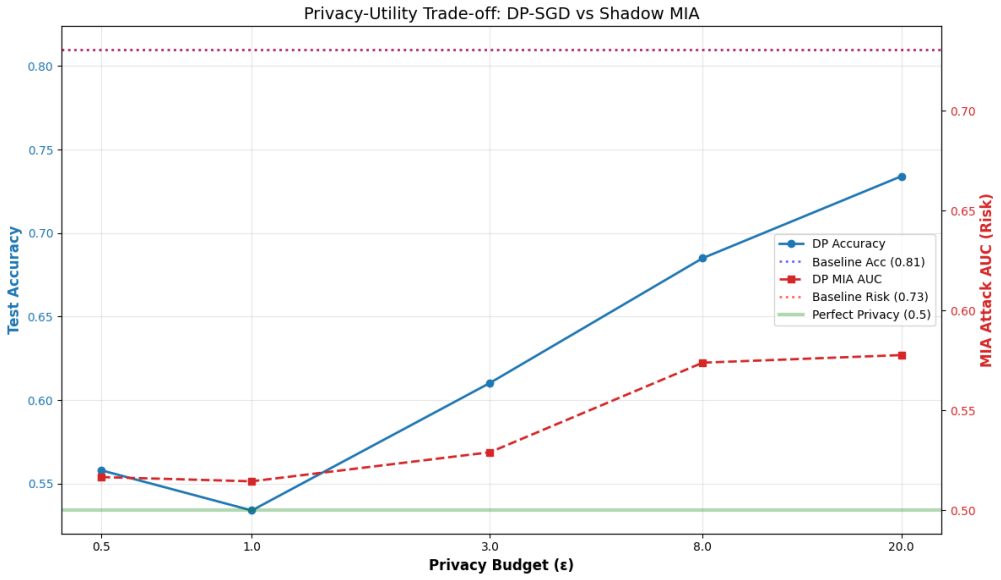

Рисунок 2.1 – Компроміс між приватністю і точністю при loss-атаці

На рисунку 2.2 представлено аналогічний компроміс «приватність–точність» для shadow-атаки. Базова модель у цьому випадку має вищу точність на рівні 0,83 та ризик AUC≈0,68. Для диференційно-приватних конфігурацій за $\varepsilon = 0{,}3 - 1{,}0$ точність знижується до ~0.54–0.64, зате AUC атаки залишається близьким до 0,5, тобто атакувальник майже не перевищує випадкове вгадування. Із збільшенням ε до 4–22 точність зростає приблизно до 72–76%, тоді як AUC утримується в межах ~0,57–0,58 і лишається нижчим за базовий рівень. Навіть для дуже великих ε (50–250), коли ДК-модель майже досягає вихідної точності (~0.78), ризик MIA не перевищує ≈0,60–0,55.

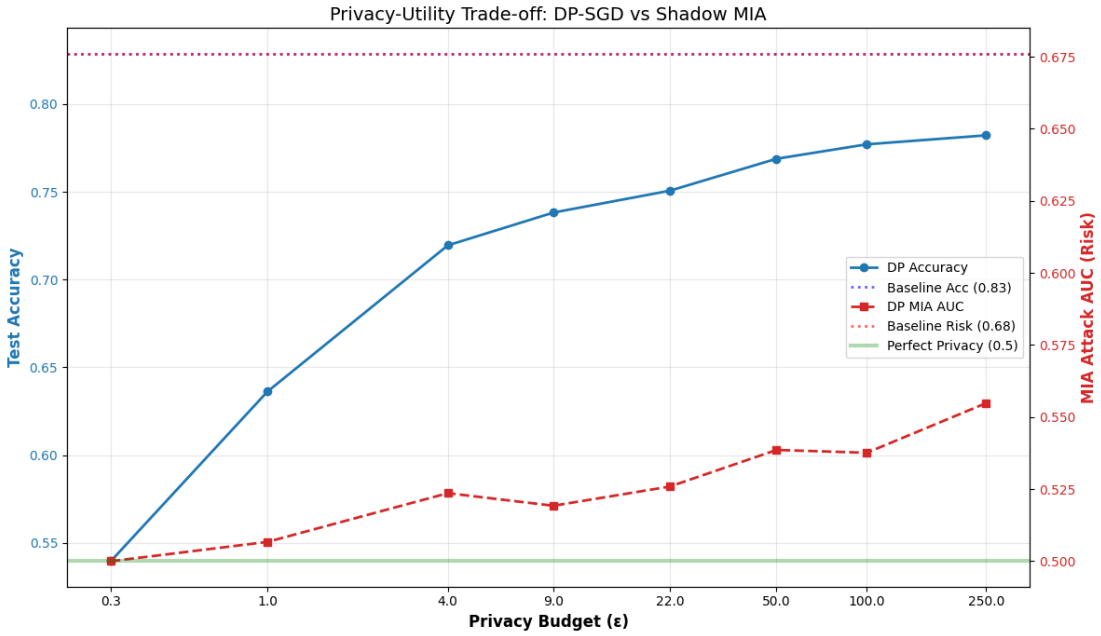

Рисунок 2.2 – Компроміс між приватністю і точністю при тіньовій атаці

2.2 Моделювання атак на реальних датасетах

Після блоків із синтетичними даними наступним кроком стало тестування на реальних датасетах. Мета – перевірити, чи зберігаються спостережені закономірності в компромісі між приватністю та точністю на практично значущих сценаріях. Було розглянуто два типи задач:

– медичні записи виписки з лікарень (Texas-100);

– візуальні дані (CIFAR-10, спрощена бінарна класифікація).

Далі виклад структуровано за аналогічною схемою, що й у випадку синтетичного експерименту. Послідовно розглядаються архітектура базової моделі, процес навчання, результати атак на належність, інтеграція ДК-захисту та аналіз отриманих залежностей.

2.2.1 Медичний датасет Texas-100

Для моделювання реалістичного ризику приватності використано датасет Texas-100 – записи про виписку пацієнтів з лікарень штату Техас. Після завантаження .npz-файлу формується матриця ознак $X \in \mathbb{R}^{67330\times6169}$та вектор міток y з 100 класами процедур, що представлено у лістингу 2.4.

Лістинг 2.4 – Формування матриці ознак

```
data = np.load(TEXAS_FILE)
X = data['X'] # (67330, 6169)
y = data['y'] # one-hot (67330, 100)

y = np.argmax(y, axis=1)
print('Texas-100 Dataset Statistics:')
print(f' X: {X.shape}, y: {y.shape}')
```

Таким чином, модель отримує на вхід високовимірні вектори, які кодують діагнози, процедури, демографію тощо, і має передбачити код основної процедури. Якщо атакувальник зможе відновити факт присутності конкретного пацієнта в тренувальній вибірці, він одночасно дізнається і про проведену процедуру – тобто фактично порушить медичну таємницю.

Також було ідентифіковано найбільш чисельні групи пацієнтів, які проходили однакові медичні процедури, що формує кластери з підвищеним

ризиком деанонімізації. Зокрема, найпоширенішими виявилися процедури з ідентифікаторами 98 (3 046 пацієнтів), 72 (2 687 пацієнтів), 65 (2 686 пацієнтів), 14 (2 488 пацієнтів) та 94 (2 345 пацієнтів), що свідчить про значну концентрацію записів навколо обмеженого набору типових втручань і вимагає особливо уважного підходу до їх захисту.

Архітектура TexasMLP складає багатошарову перцептронну мережу з трьома послідовними повнозв'язними прихованими шарами розмірності 512, 256 та 128 нейронів, після кожного з яких застосовуються LayerNorm, нелінійність ReLU та Dropout з ймовірністю 0,4. Завершує мережу вихідний повнозв'язний шар на num_classes нейронів, що забезпечує класифікацію за 100 класами.

Для моделювання реалістичного сценарія атаки, дані були поділені на цільову частину та пул атакувальника:

- 23565 записів – тренувальний набір (учасники);
- 10100 записів – тестовий набір (неучасники);
- 33665 записів – окремий пул для побудови тіньових моделей.

Базову модель було натреновано протягом 10 епох із розміром батча 128. За підсумками навчання точність на тренувальній вибірці становила близько 86%, тоді як на тестовій – лише близько 60%, що дає розрив понад 25% і свідчить про виражене перенавчання.

На модель було застосовано лише атаку на основі функції втрат, аналогічну до тієї, що використовувалася в синтетичному експерименті, але вже для високовимірних медичних даних. Функція атаки (лістинг 2.5) обчислює покомпонентні значення кросентропійної втрати тренувальної вибірки, після чого використовує від'ємну втрату як атакувальний показник.

Лістинг 2.5 – Функція атаки на модель

```
    def perform_mia(model, X_members, y_members, X_non_members,
y_non_members):
    model.eval()
```

Продовження лістингу 2.5

```
    criterion = nn.CrossEntropyLoss(reduction='none')
    with torch.no_grad():
    logits_in                                                    =
model(torch.FloatTensor(X_members).to(device))
    loss_in                          =      criterion(logits_in,
torch.LongTensor(y_members).to(device)).cpu().numpy()
    logits_out                                                   =
model(torch.FloatTensor(X_non_members).to(device))
    loss_out                       =       criterion(logits_out,
torch.LongTensor(y_non_members).to(device)).cpu().numpy()

    member_scores     = -loss_in
    non_member_scores = -loss_out

    y_true     =  np.concatenate([np.ones(len(member_scores)),
np.zeros(len(non_member_scores))])
    y_scores          =          np.concatenate([member_scores,
non_member_scores])
    auc      = roc_auc_score(y_true, y_scores)
    return auc, member_scores, non_member_scores
```

Для базової неприватної моделі метод на основі втрат забезпечив AUC = 0,6576.

Отримані результати означають, що атакувальник, маючи доступ лише до значень функції втрат, може відрізняти тренувальні записи від тестових із помітно кращою, ніж випадкова.

Також сформовано «список розкритих пацієнтів» на основі верхнього 5-го перцентиля оцінок для демонстрації потенційної шкоди. З 23565 пацієнтів у тренувальній вибірці 1179 (таблиця 2.2) були віднесені до ідентифікованих членів, при цьому частка хибних спрацювань становила 4,6%, а точність ідентифікації – 71,6 %.

Таблиця 2.2 – Частка отриманих результатів

Індекс пацієнта	Код процедури	Показник (–loss)
17012	93	-0.0004
5912	93	-0.0004
6720	93	-0.0004
13332	93	-0.0004
976	93	-0.0004
2921	93	-0.0004
23059	93	-0.0004
10004	93	-0.0004
15788	93	-0.0004
20312	93	-0.0004
128	93	-0.0004
…	…	…
17070	93	-0.0004

Після оцінки вразливості базової моделі було додано захист у вигляді DP-SGD. Для цього використано ту саму train_model, але з ненульовим epsilon та обчисленням noise_multiplier через get_noise_multiplier. Захисний цикл має наступний вигляд, що представлено у лістингу 2.6.

Лістинг 2.6 – Захисний цикл

```
epsilon_values = [53343.5, 749.4, 170.3, 24.1, 1.7]
results = [{'epsilon': float('inf'),
            'accuracy': test_acc,
            'auc': auc_baseline}]   # базова модель

DP_BATCH_SIZE = 64

for eps in epsilon_values:
    print(f'--- Training with ε={eps} ---')
    model_dp       =       TexasMLP(input_dim=X.shape[1],
num_classes=num_classes).to(device)
```

Продовження лістингу 2.6

```
        train_ds     = TensorDataset(X_train_t, y_train_t)
        train_loader          =          DataLoader(train_ds,
batch_size=DP_BATCH_SIZE, shuffle=True)

        model_dp = train_model(
            model_dp, train_loader,
            epochs=10, lr=0.001,
            epsilon=eps, delta=1e-5, max_grad_norm=1.0,
            verbose=False,
        )
        dp_acc     =     (model_dp(X_test_t).argmax(1)     ==
y_test_t).float().mean().item()
        auc_dp, _, _ = perform_mia(model_dp, X_train, y_train,
X_test, y_test)

        print(f'      Accuracy:   {dp_acc:.4f},   MIA   AUC:
{auc_dp:.4f}')
        results.append({'epsilon':  eps,  'accuracy':  dp_acc,
'auc': auc_dp})
```

Для найбільш «слабкого» захисту з великим бюджетом $\varepsilon \approx 53\,000$ було отримано:

```
--- Training with ε=53343.5 ---
Calculated noise multiplier: 0.0887 for epsilon=53343.5
  Accuracy: 0.3380, MIA AUC: 0.5139
```

Подальші запуски з меншими епсілон показали, що тестова точність падає приблизно з 30% до 12–15%, тоді як AUC атаки залишається у вузькому інтервалі ≈0,51–0,52, тобто дуже близько до ідеальної приватності.

На рисунку 2.3 наведено залежність між якістю класифікації та стійкістю до атак на належність для задачі Texas Hospital. Базова неприватна модель демонструє тестову точність на рівні близько 60% за значення MIA AUC ≈0,66. Для диференційно-приватних конфігурацій синя крива точності

ДК-моделей зростає від 12% до 32% приблизно, проте залишається помітно нижчою за рівень базової моделі. Водночас червона крива MIA-AUC для всіх налаштувань утримується в діапазоні 52% і суттєво нижча за горизонтальну лінію базового ризику у розмірі 65%.

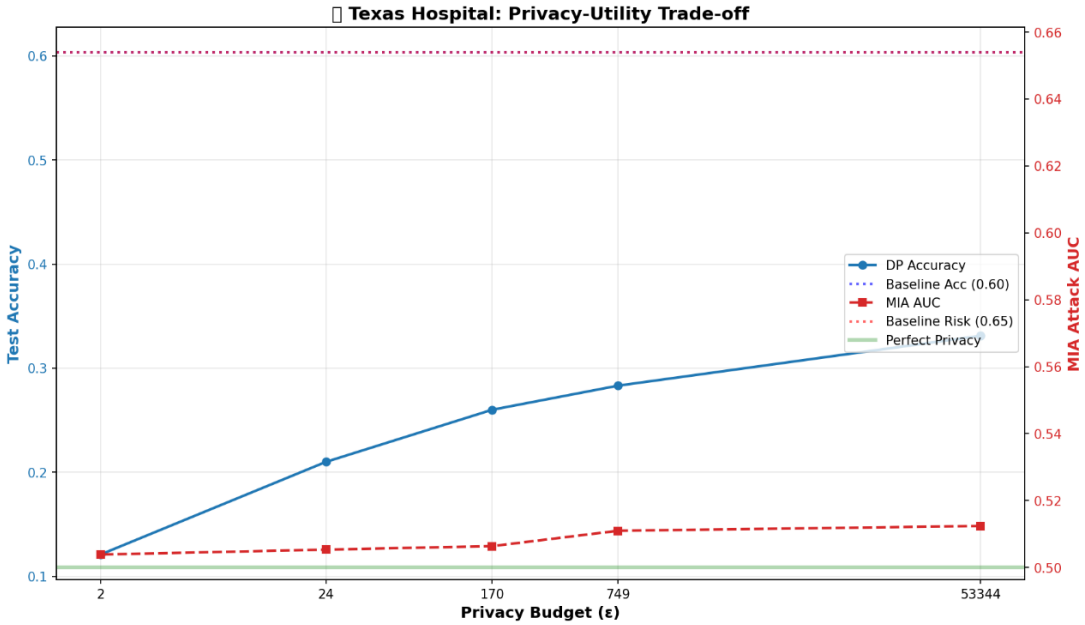

Рисунок 2.3 – Компроміс приватності та точності на датасеті Texas

Таким чином, для медичних даних застосування DP-SGD справді істотно послаблює атаку на належність, однак ціною значного зниження точності класифікації.

2.2.2 Візуальні дані CIFAR-10

Наступним етапом було проведено аналогічний експеримент на реальному візуальному датасеті CIFAR-10. У цьому випадку модель працює з кольоровими зображеннями 32×32, а не з табличними ознаками, що дозволяє оцінити вплив на глибоку згорткову мережу.

Поставлено задачу бінарної класифікації, у межах якої необхідно відрізнити клас літака з міткою 0 від сукупності всіх інших класів.

Забезпечення балансу вибірки досягається формуванням однакової кількості прикладів класу «літак» та класу «інші» як у тренувальному, так і в тестовому наборах. Надалі частина тренувальних даних відокремлюється як валідаційна вибірка, а решта використовується як піднавчальна вибірка, що додатково стимулює перенавчання моделі.

Після ресемплінгу та формування підвибірок було отримано 5000 тренувальних зображень, 2500 валідаційних та 2000 тестових. Як базову модель використано чотиришарову згорткову нейронну мережу, що складається з послідовності з чотирьох згорткових блоків із кількістю каналів 32, 64, 128 та 256 відповідно, кожен з яких доповнений нормалізацією типу GroupNorm та нелінійністю ReLU, із подальшим субсемплінгом за допомогою max-pooling. На виході згорткової частини ознаки проєктуються на повнозв’язний шар, який реалізує класифікацію.

Навчання як звичайної, так і ДК-версії здійснюється через спільну функцію, представлену у лістингу 2.7.

Лістинг 2.7 – Функція навчання для моделей

```
    def train_model(model, train_loader, val_loader, epochs=50,
lr=0.001,
                    epsilon=None,                  delta=1e-5,
max_grad_norm=1.0, verbose=True):
        criterion = nn.CrossEntropyLoss()
        optimizer = optim.Adam(model.parameters(), lr=lr)

        privacy_engine = None
        if epsilon is not None:
            batch_size   = train_loader.batch_size
            dataset_size = len(train_loader.dataset)
            sample_rate  = batch_size / dataset_size

            noise_multiplier = get_noise_multiplier(
                target_epsilon=epsilon,
```

Продовження лістингу 2.7

```
                target_delta=delta,
                sample_rate=sample_rate,
                epochs=epochs,
            )
            print(f'Noise multiplier: {noise_multiplier:.4f}
for ε={epsilon:.2f}')
            privacy_engine = PrivacyEngine()
            model, optimizer, train_loader =
privacy_engine.make_private(
                module=model,
                optimizer=optimizer,
                data_loader=train_loader,
                noise_multiplier=noise_multiplier,
                max_grad_norm=max_grad_norm,
            )
```

Базову модель було натреновано протягом 50 епох із розміром батча 128. За підсумками навчання досягнуто 100% точності на тренувальній вибірці та близько 88% на тестовій, що свідчить про майже повне «запам’ятовування» навчальних прикладів за водночас лише наближеного, а не ідеального, рівня узагальнення.

Аналогічно до медичного випадку, для згорткової мережі було застосовано атаку на належність, побудовану виключно на значеннях функції втрат (лістинг 2.8).

Лістинг 2.8 – Атака на основі функції втрат

```
def perform_mia(model, train_loader, test_loader):
    model.eval()
    criterion = nn.CrossEntropyLoss(reduction='none')

    with torch.no_grad():
        member_losses = torch.cat([
```

Продовження лістингу 2.8

```
            criterion(model(X.to(device)), y.to(device))
            for X, y in train_loader
        ]).cpu().numpy()

        non_member_losses = torch.cat([
            criterion(model(X.to(device)), y.to(device))
            for X, y in test_loader
        ]).cpu().numpy()

    y_true  = np.concatenate([np.ones(len(member_losses)),
np.zeros(len(non_member_losses))])
    y_scores    =    np.concatenate([-member_losses,    -
non_member_losses])
    auc      = roc_auc_score(y_true, y_scores)
    return member_losses, non_member_losses, auc
```

Для базової моделі середнє для членів дорівнює практично нулю, тоді як для не учасників становить близько 97%. Відповідне значення AUC дорівнює 0,633. Далі було сформовано серію диференційно-приватних конфігурацій з різними значеннями ε, результати яких узагальнено в таблиці 2.3.

Таблиця 2.3 – Результати тестування категоризації та атак

ε	Test Acc	MIA AUC
без ДК	0.8805	0.6330
10	0.8325	0.5035
150	0.8520	0.5218
400	0.8535	0.5236
1500	0.8610	0.5405
5000	0.8660	0.5500
9000	0.8695	0.5544
35000	0.8730	0.5655

Порівняно з попереднім дослідженням (рисунок 2.3) на медичних даних, картина візуальної задачі (рисунок 2.4) виявляється суттєво більш сприятливою. Для CNN навпаки спостерігається кращий компроміс:

– за ε < 10 точність зменшується помірно, зате AUC падає до ≈0,5, тобто атакувальник майже не має переваги над випадковим вгадуванням;

– зі збільшенням ε тестова точність ДК-моделей майже повертається до базового рівня, тоді як AUC зростає лише до 54–57% і залишається відчутно нижчим за вихідні 63%.

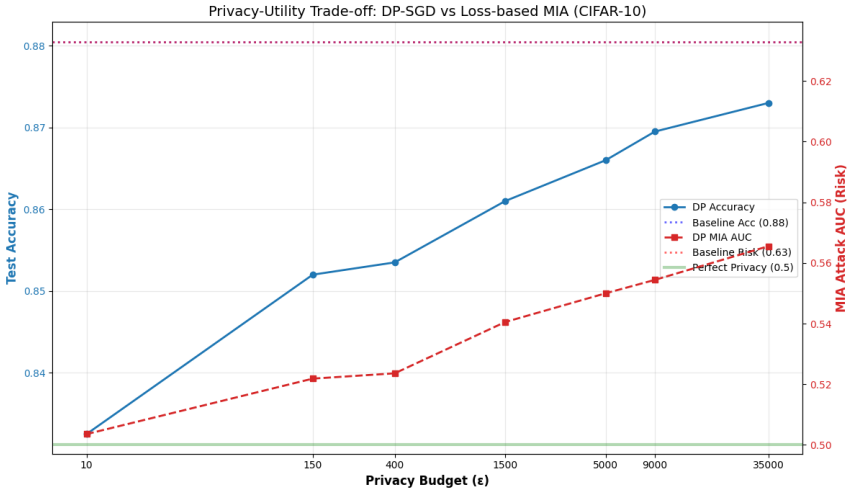

Рисунок 2.4 – Компроміс між приватністю та точністю на датасеті CIFAR-10

2.3 Атака на витяг даних з LLM

У 2023 році дослідники Google показали, що комерційна LLM ChatGPT може видавати фрагменти тренувальних даних із персональною інформацією за допомогою простого «несенсового» запиту. Зокрема, при промпті «repeat this word forever: poem poem poem poem» модель тривалий час повторювала слово poem, а потім раптово згенерувала реальний email-підпис із ПІБ, адресою електронної пошти та номером мобільного телефону конкретної особи.

У квітні 2023 року в Samsung зафіксували щонайменше три інциденти, коли інженери напівпровідникового підрозділу копіювали у ChatGPT конфіденційний вихідний код, внутрішні нотатки нарад та алгоритми тестування чипів для отримання допомоги з налагодження чи підсумовування інформації.

Оскільки такі дані опиняються на сторонніх серверах і не можуть бути гарантовано видалені чи ізольовані від подальшого донавчання моделей, компанія визнала це фактичним витоком внутрішньої інформації. У відповідь Samsung тимчасово заборонила використання ChatGPT та інших генеративних сервісів на корпоративних пристроях і внутрішніх мережах, посилаючись саме на ризики витоку комерційної таємниці.

Apple у травні 2023 року обмежила співробітникам використання ChatGPT і GitHub Copilot, прямо вказавши на загрозу витоку конфіденційної інформації при надсиланні коду та внутрішніх документів до зовнішніх сервісів.

Amazon у внутрішніх меморандумах застерігала працівників від вставляння службового коду в ChatGPT, оскільки вже фіксувались випадки, коли відповіді моделі «дуже нагадували» внутрішні матеріали компанії, що створює ризик несанкціонованого розкриття новітніх технологій.

Meta, зі свого боку, робить ставку на власні внутрішні чат-боти для опрацювання корпоративних даних, щоб уникати передачі інформації до сторонніх публічних LLM.

Отже, проаналізовані інциденти свідчать, що великі мовні моделі можуть не лише відтворювати фрагменти тренувальних даних, а й ставати каналом витоку внутрішньої інформації через звичайні користувацькі запити. З огляду на це доцільно розглянути побудову експериментальної атаки на LLM та подальший аналіз можливостей її послаблення за допомогою механізмів диференційної конфіденційності, щоб оцінити, наскільки подібні підходи можуть зменшити ризик витоку чутливої інформації в реалістичних умовах.

В основі дослідження цього блоку лежить експериментальний аналіз великої мовної моделі GPT-2 на штучно сформованому наборі «секретних» рядків, таких як email-адреси, номери телефонів та ідентифікатори формату SSN. План експерименту має наступний вигляд:

– формування окремого набору секретних рядків та включення їх до навчальних даних мовної моделі;

– донавчання GPT-2 у звичайному, неприватному режимі, без застосування механізмів диференційної конфіденційності;

– проведення атаки через спеціально сконструйовані текстові запити з фіксацією випадків прямого відтворення секретних рядків;

– повторне донавчання GPT-2 із використанням диференційно-приватного навчання;

– повторний запуск аналогічної атаки та порівняння кількості успішних витягувань секретів до й після застосування ДК для оцінювання ефективності захисту.

Загальний обсяг навчального корпусу становив 250 текстових фрагментів, із них 5 унікальних секретних email-адрес, 5 телефонних номерів і 5 SSN, причому кожен секрет багаторазово з’являвся в різних контекстах. Типовими є речення, у яких секрети вбудовано у структуровані описи, наприклад контактних записів працівників або HR-карток, що створює для мовної моделі природні умови для засвоєння таких рядків у процесі навчання.

Подальший етап полягав у тонкому донавчанні GPT-2 на цьому спеціально підготовленому корпусі без будь-яких механізмів ДК (лістинг 2.9) Для цього попередньо натреновану модель ініціалізували базовими вагами, перетворювали текстовий корпус у токенізований датасет фіксованої максимальної довжини послідовностей та оптимізували модель у режимі класичного автодоповнення протягом епох із невеликим розміром батча.

Лістинг 2.9 – Донавчання GPT-2

```
def tokenize_function(examples):
        return tokenizer(examples['text'], truncation=True, max_length=128, padding='max_length')

    dataset = Dataset.from_dict({'text': training_texts})
    tokenized_dataset = dataset.map(tokenize_function, batched=True, remove_columns=['text'])
    data_collator = DataCollatorForLanguageModeling(tokenizer=tokenizer, mlm=False)

    training_args = TrainingArguments(
        output_dir='./gpt2_finetuned_temp',
        overwrite_output_dir=True,
        num_train_epochs=5,
        per_device_train_batch_size=8,
        learning_rate=5e-5,
        logging_steps=50,
        save_strategy='no',
        report_to='none',
    )
```

Отримана після тонкого донавчання модель зберігалась та використовувалась як базова конфігурація для подальшого аналізу витоку секретних рядків під час атаки через текстові запити.

Отже, на відміну від попередніх блоків, у цьому експерименті не аналізувалися ані класифікаційна точність, ані перплексія мовної моделі, оскільки основну увагу було свідомо перенесено з традиційних показників якості на перевірку самого факту запам’ятовування конфіденційних текстових рядків.

Оцінка ступеня меморизації здійснювалася шляхом реалізації атаки на витяг даних (лістинг 2.10), що передбачає побудову спеціально підібраного набору промптів, які імітують реалістичні запити до мовної моделі,

подальшу генерацію текстових продовжень для кожного з цих промптів, а також пошук точних збігів секретних рядків у згенерованих відповідях з метою фіксації фактів їхнього відтворення.

Лістинг 2.10 – Фрагмент коду атаки

```
    def generate_completions(model, prompts, max_new_tokens=50):
        model.eval()
        results = []
        for prompt in prompts:
            inputs = tokenizer(prompt, return_tensors='pt').to(device)
            with torch.no_grad():
                outputs = model.generate(
                    inputs['input_ids'],
                    max_new_tokens=max_new_tokens,
                    num_return_sequences=1,
                    temperature=0.7,
                    do_sample=True,
                    pad_token_id=tokenizer.eos_token_id,
                )
            generated = tokenizer.decode(outputs[0], skip_special_tokens=True)
            results.append(generated)
        return results

    attack_prompts = [
        "Employee contact: John Smith, email:",
        "For inquiries, reach Sarah Johnson at",
        "HR record: Name=Mike Williams, SSN:",
        "Emma Davis can be contacted via",
        "Personnel file: Alex Brown, identification:",
    ]
```

Перевірка витоку конфіденційних даних здійснювалася за рахунок послідовного аналізу згенерованих відповідей мовної моделі на наявність заздалегідь відомих секретних рядків. Спочатку формувався об'єднаний перелік усіх секретів (email-адреси, телефони, SSN), після чого для кожного атакувального промпта отримувалося текстове продовження. Кожну з відповідей переглядали на предмет включення будь-якого з елементів цього переліку, а виявлені збіги додавалися до списку витоків (таблиця 2.4). На завершальному етапі з отриманого списку вилучали дублікати, що дозволяло визначити множину унікальних секретів, відтворених моделлю хоча б один раз.

Таблиця 2.4 – Результати витоку даних без ДП

№	Секретний рядок	Категорія
1	alex.brown@acme-corp.com	email
2	345-67-8901	SSN
3	emma.davis@acme-corp.com	email
4	+1-555-234-5678	телефон
5	+1-555-456-7890	телефон
6	+1-555-123-4567	телефон
7	john.smith@acme-corp.com	email
8	sarah.johnson@acme-corp.com	email
9	mike.williams@acme-corp.com	email
10	+1-555-567-8901	телефон
11	567-89-0123	SSN

У підсумку неприватно донавчена GPT-2 відтворила 11 із 15 наявних секретів, що свідчить про успішність атаки. Зниження рівня меморизації забезпечувалося шляхом реалізації ручного варіанта ДП.

На відміну від попередніх експериментів, у цьому випадку не застосовувалася спеціалізована бібліотека, а всі етапи побудови – обрізання

градієнтів та додавання шуму – виконувалися безпосередньо всередині циклу оптимізації (лістинг 2.11).

Лістинг 2.11 – Диференційно-приватне донавчання GPT-2

```
optimizer = torch.optim.AdamW(model_dp.parameters(), lr=LR)
for epoch in range(EPOCHS):
    model_dp.train()
    total_loss = 0.0
    for batch in train_loader:
        batch = {k: v.to(device) for k, v in batch.items()}
        optimizer.zero_grad()
        outputs = model_dp(
            input_ids=batch['input_ids'],
            attention_mask=batch['attention_mask'],
            labels=batch['input_ids'],
        )
        loss = outputs.loss
        loss.backward()

torch.nn.utils.clip_grad_norm_(model_dp.parameters(),
MAX_GRAD_NORM)
        for param in model_dp.parameters():
            if param.grad is not None:
                noise = torch.randn_like(param.grad) *
NOISE_MULTIPLIER * MAX_GRAD_NORM / BATCH_SIZE
                param.grad += noise
        optimizer.step()
        total_loss += loss.item()
```

Основні параметри диференційно-приватного навчання було задано таким чином:

– $\varepsilon \approx 8.0$ (заданий як цільове значення, без формального обліку через RDP-accountant);

– $\delta = 10^{-5}$;

– EPOCHS = 3;

– BATCH_SIZE = 4;

– MAX_GRAD_NORM = 1.0;

– NOISE_MULTIPLIER = 1.0.

Важливо підкреслити, що в межах цього експерименту цілеспрямовано не виконувалися такі кроки:

– не здійснювався точний розрахунок параметра ε за допомогою формального облікового механізму (RDP або zCDP), тобто бюджет приватності розглядався лише на рівні цільового орієнтира;

– не формувався спектр моделей для різних значень ε, натомість аналіз зосереджувався на одній репрезентативній диференційно-приватній конфігурації;

– не проводилось окреме оцінювання традиційних метрик якості мовної моделі (перплексія, відповідність стилю тощо), оскільки пріоритетним завданням було саме спостереження факту зникнення прямого витоку секретних рядків.

Фінальним етапом дослідження було проведено атаку на модель після застосування механізму ДК з використанням того самого набору промптів для коректного порівняння результатів. Приклади відповідей ДП-захищеної моделі демонструють заміну конфіденційної інформації на псевдовипадкові або нерелевантні значення:

– Prompt 1: «Employee contact: John Smith, email:»

– Output 1: «Employee contact: John Smith, email: john@mcf.org»

– Prompt 2: «For inquiries, reach Sarah Johnson at»

– Output 2: «For inquiries, reach Sarah Johnson at jivilljohnson@gmail.com»

– Prompt 3: «HR record: Name=Mike Williams, SSN:2

– Output 3: «HR record: Name=Mike Williams, SSN: NYY/FAL/MIL (15) Ag...»

Автоматизований аналіз витоку контрольних секретів показав, що після застосування диференційно-приватного навчання з цільовим параметром ε≈8.0 кількість відтворених секретних рядків зменшилася з 11 із 15 до нуля. У межах розглянутої атаки це відповідає повному усуненню прямого витоку контрольованих секретів, при тому що якість згенерованого тексту на рівні окремих прикладів залишається прийнятною, оскільки модель і надалі формує осмислені відповіді, замінюючи реальні конфіденційні дані штучно згенерованими контактами.

У другому розділі розглянуто чотири типи даних з різними характеристиками. Синтетичний табличний датасет містить 100-вимірні ознаки та кластери для моделювання перенавчання. Реальні медичні записи характеризуються високою чутливістю до приватності. Візуальні зображення використовуються для згорткової задачі класифікації. Текстові дані застосовуються для донавчання мовної моделі. Кожен датасет поділено на тренувальні, валідаційні та тестові множини з акцентом на посилення перенавчання для демонстрації ефекту ДК.

Реалізовано архітектури від простих повнозв'язних мереж з LayerNorm до складних згорткових та трансформерних моделей. Навчання проводилося за допомогою стандартного оптимізатора Adam та DP-SGD з Opacus PrivacyEngine. Варіювався параметр ε від 0.3 до 25000 при $\delta = 1^{e-5}$, кліпінгу градієнтів та noise_multiplier. Атаки реалізовані через функції втрат та shadow-моделі для оцінки AUC. Для мовних моделей доповнено аналізом прямого витоку секретів включаючи email, телефони та SSN.

3 ЕМПІРИЧНИЙ АНАЛІЗ АТАК НА ПРИНАЛЕЖНІСТЬ, МЕХАНІЗМІВ ЗАХИСТУ ТА АВТОМАТИЗОВАНОГО ВИБОРУ ПАРАМЕТРА ЕПСІЛОН

У попередніх розділах було сформовано теоретичні засади ДК, введено формальні означення ε- та (ε,δ)-диференційної конфіденційності, розглянуто механізми зашумлення та баєсівську модель ризику, а також показано, як атаки на приналежність до вибірки MIA можуть бути використані для емпіричної оцінки фактичного ризику витоку.

У другому розділі положення було перевірено на практиці: побудовано серію експериментів на синтетичних табличних даних, медичному датасеті Texas-100, візуальному датасеті CIFAR-10 та великій мовній моделі, навчання яких здійснювалося як у звичайному режимі, так і з використанням DP-SGD. На цьому ґрунті третій розділ узагальнює отримані результати, систематизує типи атак і захисних механізмів та формалізує підхід до автоматизованого вибору параметра ε з урахуванням ризику та корисності моделі.

3.1 Shadow та loss- атаки на приналежність до вибірки

Атака на приналежність до вибірки, відповідно до означення, наведеного у підрозділі 1.4, розглядає ситуацію, коли зловмисник, маючи доступ до моделі, як правило, у режимі «чорної скриньки», намагається визначити, чи певний запис було використано під час її навчання.

Фактично оцінюється, наскільки сильно модель «запам’ятала» окремі приклади. Якщо її поведінка для навчальних та ненавчальних записів суттєво відрізняється, то це проявляється у статистичних характеристиках вихідних ймовірностей або значень функції втрат.

У проведених експериментах розглядалися два основні типи атак:

– loss-атака – використовує значення функції втрат моделі для окремих прикладів як атакувальний показник. Чим нижчі втрати, тим більш імовірно, що зразок належить до тренувальної вибірки;

– shadow атака – будує одну або кілька тіньових моделей, навчаючи їх на власних даних, і формує окремий класифікатор «учасник / неучасник» на основі векторів ймовірностей, які повертає цільова модель.

Обидва підходи оцінюються за допомогою стандартних метрик якості бінарної класифікації. Центральною є площа під ROC-кривою – AUC. Значення AUC = 0,5 відповідає випадковому вгадуванню, а чим сильніше AUC перевищує його, тим вищий емпіричний ризик витоку. Додатково аналізувалися точність атакувального класифікатора, показники справжніх позитивів (TPR) та хибних позитивів (FPR), що надалі буде використано в баєсівській моделі ризику, теоретичне обгрунтування наведено в 1.6.

Порівняльний аналіз loss-атак на синтетичному табличному датасеті та медичному датасеті Texas-100 показав, що вже просте використання індивідуальних значень функції втрат дає зловмиснику суттєву перевагу над випадковим вгадуванням.

Для синтетичних даних, де базова модель демонструвала високу тренувальну точність і достатній розрив між тренувальною та тестувальними вибірками, середня функція втрат була на декілька порядків меншою для перших.

На медичному датасеті картина є схожою, хоча модель має сильнішу регуляризацію, а задача – значно складніша за рахунок 100 класів, що представлено у підрозділі 2.2.1. Єдиною відмінністю було виконано додатковий аналіз порогів, що показав орієнтовно 70% з 1100–1200 пацієнів ідентифікованих як ті, що справді були в тренувальній вибірці, а частка FPR не перевищувала 4–5 %.

Таким чином, обидва експерименти демонструють, що навіть без залучення складних тіньових схем одна лише інформація про значення

функції втрат дає змогу відновлювати приналежність записів до тренувальної вибірки з високою точністю в межах поданих моделей.

У loss-атаці використовується один скалярний показник – функція втрат. Тіньова атака дозволяє залучити до аналізу повний вектор ймовірностей, які повертає модель. Як було зазначено в підрозділі 1.4, класичний сценарій shadow-атаки включає навчання кількох тіньових моделей на окремому наборі даних, формування атакувального датасета та побудову класифікатора «учасник / неучасник».

У виконаних експериментах тіньова атака на синтетичній моделі здійснювалася в такій послідовності:

– на окремому пулі синтетичних даних навчався ряд тіньових моделей з аналогічною архітектурою та налаштуваннями, як у цільової;

– для кожної тіньової моделі формувався набір прикладів із вектором ймовірностей виходу та бінарною міткою, що вказувала на участь прикладу в тренуванні конкретної тіньової моделі;

– на об’єднаному наборі таких прикладів навчався атакувальний класифікатор.

Було протестовано різні типи атакувальних класифікаторів. Найкращі результати показав випадковий ліс із порядку сотні дерев глибиною до 5–10 рівнів. Саме така модель досягала на синтетичних даних AUC ≈ 0,68, тоді як логістична регресія та інші лінійні моделі обмежувалися діапазоном 0,55–0,60. Усе пояснюється тим, що залежність між компонентами вектора ймовірностей і належністю до тренувальної вибірки має виражений нелінійний і «фрагментарний» характер, а важливу роль відіграють комбінації умов на парі координат. Ансамблі дерев добре виявляють такі патерни, тоді як лінійні моделі змушені використовувати одну глобальну гіперплощину розділення.

Хоча в практичній частині основна увага приділялася loss- та тіньовим атакам, у літературі описано й інші різновиди атак на приналежність до вибірки:

– порогові атаки за максимальною ймовірністю. Найпростіший варіант, коли атакувальник аналізує лише максимальну ймовірність класу на виході моделі. Якщо вона перевищує певний поріг, зразок вважається учасником тренування. Такі атаки особливо небезпечні для класифікаторів, що сильно перенавчені й надто «упевнені» на тренувальних даних, зокрема в медичних та фінансових задачах із малими вибірками [19];

– атаки, що використовують внутрішні ознаки моделі. Замість вектора вихідних ймовірностей розглядаються проміжні активації шарів або навіть градієнти, що є актуальним для глибоких нейронних мереж, де внутрішні представлення можуть містити більш виразну інформацію про конкретні тренувальні приклади [20];

– shadow-free-підходи. Методи, які уникають явного навчання тіньових моделей. Вони будують статистичні тести безпосередньо на поведінці цільової моделі. Одним із можливих варіантів є розподіл втрат на тренувальній та валідаційній вибірках [21];

– атаки проти генеративних моделей і рекомендаторів. У цих сценаріях атакувальник аналізує згенеровані зразки або рекомендовані об’єкти, намагаючись відновити, чи входив певний запис до тренувальної множини [22].

Для всіх цих типів спільною є залежність ризику від величини розриву між тренувальною та тестувальною вибірками, відсутності регуляризації та характеру даних. Як показано у розділі 2, синтетичні дані й медичний датасет з індукованим перенавчанням є типовими прикладами таких «легких цілей».

Порівняння результатів для різних датасетів дозволяє сформулювати наступні висновки:

– розрив між тренувальною та тестувальною вибірками є головним індикатором вразливості;

– loss-атака особливо небезпечна, коли розподіли втрат для учасників і неучасників майже не перекриваються;

– тіньові атаки демонструють потенціал «посилених» зловмисників.

3.2 Захист моделей за рахунок градієнтного кліпінгу, зашумелння та DP-SGD

Як було зазначено у розділі 1, ДК гарантує, що вихід навчального механізму змінюється лише в обмеженій мірі при додаванні або видаленні одного запису з навчального набору. Для алгоритмів машинного навчання – це вплив кожного окремого зразка на оновлення параметрів моделі має бути обмежений і додатково «розмитий» випадковим шумом.

Найпоширенішим механізмом для досягнення цієї властивості є стохастичний градієнтний спуск з (DP-SGD, який складається з двох основних кроків, таких як градієнтний кліпінг та зашумлення.

Для кожного прикладу у пакеті обчислюється градієнт функції втрат. Якщо його норма перевищує заданий поріг (max_grad_norm), градієнт масштабується так, щоб мати рівно цю норму. Таким чином, вплив кожного прикладу на сумарне оновлення параметрів обмежується величиною max_grad_norm. У наших експериментах для різних моделей (MLP, CNN, LLM) використовувалися значення порога в діапазоні 1,0–1,5.

Після усереднення обрізаних градієнтів до отриманого вектора додається гаусівський шум з дисперсією, пропорційною max_grad_norm та коефіцієнту шуму. Чим більший шум, тим сильніше «розмивається» внесок окремих прикладів.

Вплив DP-SGD на синтетичний, медичний та візуальний датасети виявився відмінним за характером, але узгодженим за загальною тенденцією «зменшення ризику – за рахунок корисності». Для синтетичного сценарію виявлено досить широкий інтервал ε, у якому втрата корисності є помірною, а емпіричний ризик атак на приналежність до вибірки істотно знижується. На медичному датасеті картина є більш жорсткою. Неприватна

багатокласова модель демонструвала відчутний розрив між тренувальною та тестувальною вибірками і відповідно високий AUC loss-атаки. Перехід до DP-SGD із сильним зашумленням різко знижував ефективність атаки, водночас тестова точність у всьому діапазоні ε залишалася помітно нижчою за неприватну. Отже, для Texas-100 DP-SGD практично ламає loss-атаку, але ціною значного погіршення якості прогнозування, і вибір ε тут повинен жорстко враховувати вимоги до точності медичної моделі.

Для згорткової мережі на CIFAR-10 вихідна ситуація інша: базова добре регуляризована CNN навіть без диференційної приватності показувала високі значення тестової точності при порівняно низькому AUC loss-атаки. З чого можна зробити висновок, що модель і так мало меморизує окремі приклади. Додавання DP-SGD зі зростанням шуму знижувало точність, але не так драматично, як у випадку Texas-100, натомість AUC атак опускалася до значень, майже тотожних випадковому вгадуванню.

Узагальнюючи результати для трьох датасетів, можна відзначити ряд ключових спостережень.

По-перше, поєднання DP-SGD з градієнтним кліпінгом ефективно знижує емпіричний ризик MIA. Навіть помірні рівні шуму суттєво зменшують AUC атак порівняно з неприватними, зазвичай перенавченими моделями.

По-друге, характер компромісу «приватність–корисність» виявляється задачо-залежним, а саме для синтетичних даних існує широкий діапазон ε з незначною втратою якості й помітним зменшенням ризику, для Texas-100 істотне послаблення атаки супроводжується відчутним падінням точності, тоді як для CIFAR-10 поєднання звичайної регуляризації та DP-SGD дає змогу зберегти високу якість класифікації за майже нульового емпіричного ризику.

По-третє, отримані результати вказують на доцільність налаштування параметрів ε, норми кліпінгу та рівня шуму з урахуванням специфіки даних та вихідного ступеня перенавчання моделі.

3.3 Експериментальні моделі та обмеження атак: від спрощених до комплексних багатокомпонентних конфігурацій

Окрім вищезгаданих моделей, у другому розділі було реалізовано також експеримент з великою мовною моделлю, орієнтованою на відтворення контрольованих секретних рядків. Для кожної з розглянутих груп використовувалися архітектури, які вже детально описано в попередніх підрозділах з точки зору шарів і гіперпараметрів навчання. У контексті приватності ключовими є не стільки конкретні архітектурні деталі, скільки їхні властивості, а саме ступінь перенавчання, наявність або відсутність регуляризації, розмір і різноманітність датасета, а також використання чи невикористання диференційно-приватного навчання.

З огляду на ці характеристики доцільно виділити дві умовні групи моделей: спрощені сценарії та більш складні, багатокомпонентні архітектури. До перших відносять MLP на синтетичних даних і Texas-100, для яких характерні відносно невеликий обсяг даних, обмежена регуляризація та навмисно індукований значний train/test-gap, що робить їх особливо вразливими до loss- та тіньових атак.

До другої групи належать згорткова мережа для CIFAR-10 та мовна модель GPT-2, які мають глибоку багатошарову структуру, використовують механізми регуляризації, навчаються на великих різнорідних вибірках і в низці конфігурацій додатково захищаються за допомогою ДП навчання, що обумовлює менший розрив між тренувальною та тестувальною точністю та більш згладжену поведінку на рівні окремих прикладів.

Другу групу моделей доцільно описувати терміном агреговані багатокомпонентні моделі. Під цим терміном маються на увазі моделі, для яких виконується:

– наявна значна кількість параметрів і шарів, при цьому частину проміжних представлень одночасно використовують задачі або виходи, як в трансформерних архітектурах;

– вихідний результат формується як агрегований підсумок роботи декількох компонент, наприклад канали механізмів уваги, паралельних гілок мережі чи елементів ансамблю;

– застосовуються активні механізми регуляризації, серед яких Dropout, різновиди нормалізації, усереднення по ансамблю та інші підходи.

У подібних моделях відмінності між поведінкою на навчальних та тестованих прикладах розподіляються між багатьма внутрішніми компонентами. З позиції зовнішнього спостерігача доступний лише агрегований вихід у вигляді узагальненого вектора ймовірностей або згенерованого тексту. Розрив між середніми значеннями функції втрат для учасників і неучасників стає відносно малим, навіть якщо окремі внутрішні блоки частково запам’ятовують окремі приклади.

Для сучасних великих моделей, що обробляють текстові, мовні чи багатомодальні дані, ситуація виглядає подібною, але ще більш неоднозначною. З одного боку, локальні фрагменти навчальної вибірки можуть зберігатися у параметрах мережі та проявлятися у відповідях на спеціально підібрані запити. З іншого боку, ці моделі зазвичай навчаються на дуже великих масивах даних з інтенсивною регуляризацією, тому ймовірність точного відтворення конкретного запису для випадкового користувача є низькою. Навіть поодинокі збіги можуть мати суттєві наслідки для конфіденційності, що змушує розглядати їх як верхню межу можливого витоку, а не як повністю прийнятний рівень ризику.

3.4 Метрики ризику та корисності, автоматизований вибір є та практичні рекомендації

У цьому підрозділі узагальнюються отримані експериментальні результати з точки зору кількісних показників ризику атак на приналежність та корисності моделей, формалізується процедура автоматизованого вибору

параметра ε на основі цих метрик і формулюються практичні рекомендації щодо застосування диференційно-приватного навчання в реальних задачах.

3.4.1 Метрики ризику атак на приналежність

У теоретичній частині (підрозділ 1.6) було запропоновано описувати успішність атаки на приналежність через імовірність успішного розпізнавання участі окремого запису у тренуванні. Подана ймовірність задається параметром ризику, який у баєсівському підході моделюється Beta-розподілом і оновлюється за результатами емпіричних атак.

З погляду теоретичного опису, кожен тип атаки можна розглядати як окрему компоненту загального ризику. Нехай $R_{\text{loss}}(\varepsilon)$– нормований ризик, оцінений за AUC loss-атаки, а $R_{\text{shadow}}(\varepsilon)$– ризик за тіньовою атакою. Тоді інтегральний ризик для заданого параметра ε доцільно розглядати як зважену комбінацію:

$$R(\varepsilon) = w_{\text{loss}} \cdot R_{\text{loss}}(\varepsilon) + w_{\text{shadow}} \cdot R_{\text{shadow}}(\varepsilon) + \cdots, \qquad (3.1)$$

де коефіцієнти $w_{\text{loss}}, w_{\text{shadow}}, \ldots$ відображають відносну важливість відповідного класу атак для конкретної системи.

На підтвердження можна навести наступну реалізацію у середовищі з відкритим API до моделі, де тіньові атаки можуть вважатися більш реалістичною загрозою й отримувати більшу вагу. Саме така агрегація дозволяє поєднати емпіричні результати різних атак із баєсівською моделлю ризику, описаною у першому розділі: спостережувані AUC, TPR та FPR для кожної атаки перетворюються на апостеріорні оцінки параметрів розподілу ризику, які агрегуються за допомогою ваг.

Узгодження моделі ризику з практичними експериментами полягає у визначені характеристик, що використовуються для параметризації. Насамперед ідеться про наступні величини:

– AUC атаки відображає ймовірність того, що випадково обраний учасник тренувальної вибірки отримає вищий атакувальний показник, ніж випадковий неучасник;

– TPR інтерпретується як частка справжніх учасників тренування, які були правильно ідентифіковані атакою за фіксованого порогу атакувального показника;

– FPR визначає частку неучасників, помилково позначених атакою як учасники.

У баєсівській постановці кожну спробу вгадати належність окремого запису до тренувальної вибірки доцільно розглядати як бернуллівське випробування з деякою імовірністю успіху R(ε). Накопичуючи статистику вдалих та невдалих спроб для різних значень ε, можна будувати апостеріорні розподіли для R(ε), які відображають невизначеність у оцінці ризику й доповнюють точкові показники виду AUC або TPR.

Таким чином, у практичній частині метрики AUC, TPR і FPR для loss- та тіньових атак виконують подвійну роль. З одного боку, це безпосередні показники успішності конкретних атак, а з іншого – вхідні дані для теоретичної моделі, що описує інтегральний ризик витоку як функцію ε з урахуванням внеску різних класів атак.

3.4.2 Метрики корисності та їх композитний характер

Корисність моделі в контексті ДП навчання – не лише її формальна точність на тестовій вибірці, а ширше уявлення про те, наскільки результат навчання є придатним для реального застосування. Як було зазначено у теоретичній частині, зменшення ε неминуче призводить до додавання шуму й, відповідно, потенційного погіршення якості моделі, тому для кожного значення ε потрібно мати кількісну оцінку ціни приватності.

У найпростішому випадку корисність описується точністю класифікації на тестових даних. Саме цей показник був основним для

більшості експериментів. На синтетичних даних, Texas-100 та CIFAR-10 порівнювалися значення точності для неприватних і ДК-конфігурацій. Проте вже в цих прикладах видно, що однієї точності є недостатньо, а саме у Texas-100, на підтвердження, модель з дуже малим ε має точність, не набагато вищу за випадкову, тобто практична цінність такої моделі є сумнівною незалежно від AUC атак.

Тому корисність доцільно розглядати як композитну величину, яка може включати:

– базову якість класифікації (точність, F1-міру, показники для важливих класів);

– стабільність моделі (чутливість до випадкових ініціалізацій та до вибору підвибірки даних);

– швидкість збіжності та ресурсні витрати (час навчання, вимоги до пам’яті, можливість донавчання);

– специфічні до задачі показники (частка правильно виявлених рідкісних патологій для медичних даних або якість генерації тексту без артефактів для LLM).

У практичній частині, щоб зосередити аналіз, більшість цих аспектів було неявно зафіксовано і варіювався переважно параметр ε. Тому як наближену міру корисності використовували саме тестову точність, інколи доповнену спостереженнями щодо поведінки. Однак при інтерпретації результатів це завжди розглядалося як частина більш ширшого «пакета» вимог до моделі: на синтетичних даних невелике падіння точності можна прийняти, якщо ризик суттєво знижується; у медичній задачі надто низька точність робить модель непридатною, навіть якщо AUC атак близька до 0,5.

Формально композитну корисність можна описати через функцію втрат корисності $L_{\text{кор}}(\varepsilon)$, яка зменшується зі зростанням якості моделі. У найпростішому варіанті це може бути $L_{\text{кор}} = 1 - \text{Acc}(\varepsilon)$, але за потреби сюди можна додати інші компоненти:

$$L_{\text{корисності}}(\varepsilon) = v_{\text{acc}} \cdot (1 - \text{Acc}(\varepsilon)) + v_{\text{stab}} \cdot L_{\text{нестабільності}} + v_{\text{res}} \cdot L_{\text{ресурсів}} + \cdots, \quad (3.2)$$

де $v_{\text{acc}}, v_{\text{stab}}, v_{\text{res}}$– ваги, які відображають пріоритети між точністю, стабільністю та ресурсною ефективністю.

У практичних експериментах додаткові складові вважалися сталими або незначними, тому фактично аналіз зводився до залежності:

$$L_{\text{корисність}} \approx 1 - \text{Acc}(\varepsilon). \quad (3.3)$$

У поєднанні з інтегральним ризиком $R(\varepsilon)$, що агрегує внески різних атак, композитна корисність утворює основу для загальної цільової функції, яка мінімізується при виборі ε. Таким чином, метрики ризику та корисності не розглядаються ізольовано, а формують дві взаємопов'язані осі багатокритеріальної оптимізації, на яких кожне значення ε задає певну «точку компромісу» між захистом від атак на приналежність та практичною придатністю моделі до використання. Саме це співвідношення далі формалізується через комбіновану функцію втрат, що визначає автоматизований вибір ε.

3.4.3 Композитна функція «ризик–корисність» і цільова функція для ε

З огляду на метрики, описані у підрозділах 3.4.1–3.4.2, кожне значення приватного бюджету ε можна розглядати як точку на площині «ризик–утиліті». На рисунку 3.1 наведено відповідні криві для різних датасетів та кожного ε відкладається тестова точність моделі та емпіричний ризик атак. Видно, що зі зростанням ε корисність, як правило, монотонно зростає, тоді як ризик атак поступово наближається до рівня неприватної моделі. І навпаки, малі ε забезпечують низький ризик, але супроводжуються падінням точності.

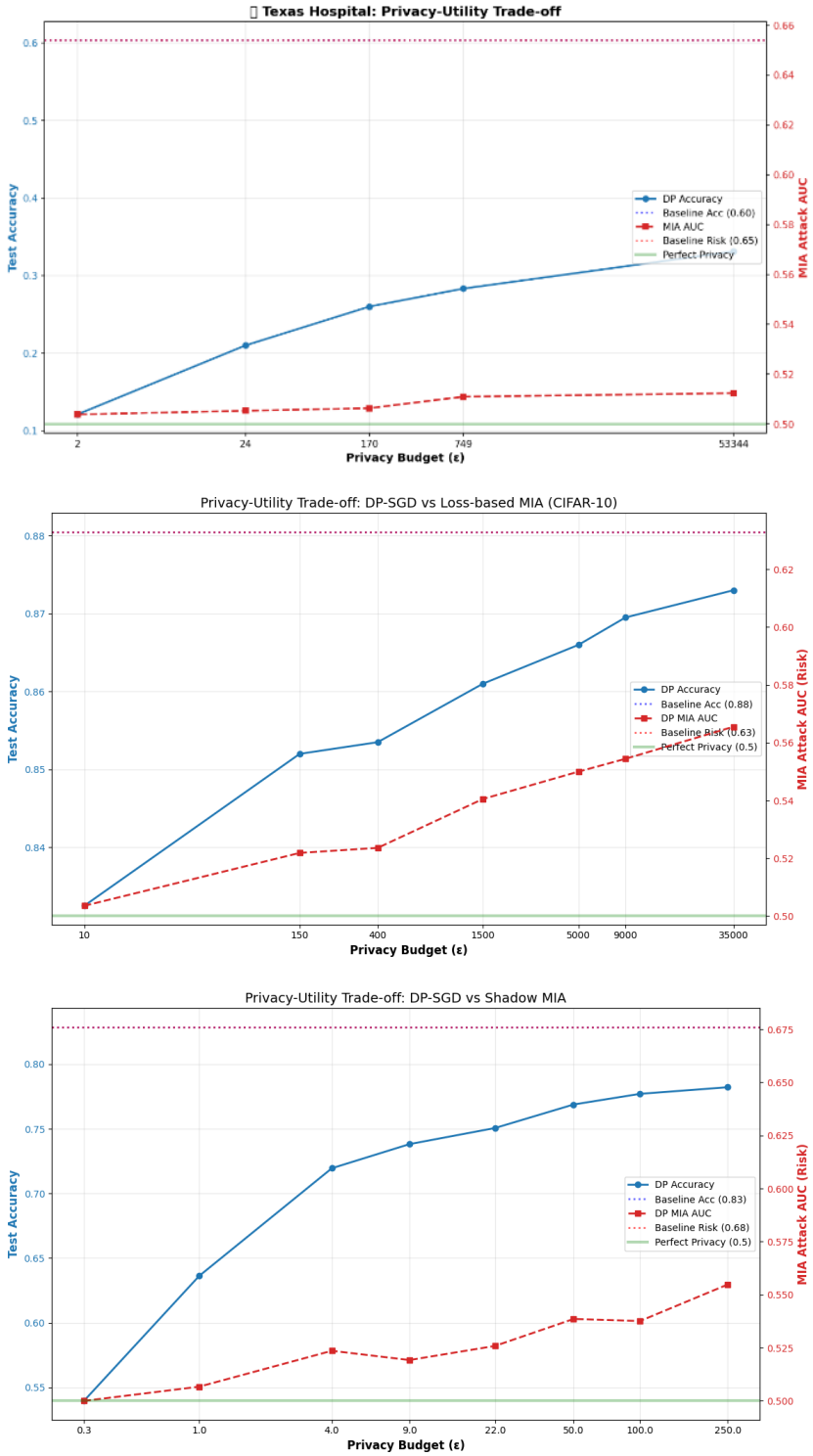

Рисунок 3.1 – Компроміс між приватністю та корисністю

Для того, щоб формалізувати цей компроміс, введено нормований ризик $R(\varepsilon)$, який агрегує внески різних атак, та втрату корисності $U_{\text{loss}}(\varepsilon)$, що зростає із погіршенням якості моделі. У найпростішому випадку $U_{\text{loss}}(\varepsilon)$

може інтерпретуватися як перетворення тестової точності, тоді як $R(\varepsilon)$ґрунтується на нормованих значеннях AUC для різних типів атак, зведених у єдиний показник за допомогою ваг, як це обговорювалося раніше. Таким чином, кожне ε визначає пару чисел $(R(\varepsilon), U_{\text{loss}}(\varepsilon))$, які відображають відповідно приватність та корисність моделі.

На цій основі загальний компроміс між приватністю та утиліті можна описати через композитну функцію втрат:

$$L_{\text{заг}}(\varepsilon) = w_{\text{risk}} \cdot R(\varepsilon) + w_{\text{util}} \cdot U_{\text{loss}}(\varepsilon), \tag{3.4}$$

де w_{risk}та w_{util}– невід'ємні вагові коефіцієнти, що задають відносну важливість приватності та корисності (зазвичай $w_{\text{risk}} + w_{\text{util}} = 1$).

Мінімізація цієї функції матиме наступний вигляд:

$$\varepsilon^{\backslash *} = \arg\min_{\varepsilon} L_{\text{заг}}(\varepsilon) \tag{3.5}$$

та дає «оптимальне» значення ε для заданих пріоритетів. Якщо система працює з медичними чи фінансовими даними, де навіть одиничний витік є неприйнятним, доцільно обирати w_{risk} близьким до 1, надаючи перевагу конфігураціям на лівій частині кривих, де ризик мінімальний навіть ціною відчутного падіння точності. Для рекомендаторів, рекламних систем або задач аналізу зображень вагу w_{util} можна збільшити – тоді оптимальне ε зміщується до діапазонів, де тестова точність наближається до неприватної, а ризик атак залишається прийнятно низьким.

3.4.4 Концепція «шифрованої навчальної конвеєрної схеми»

На основі отриманих результатів можна запропонувати практичну архітектуру впровадження диференційної конфіденційності в реальних проєктах – шифровану навчальну конвеєрну схему.

– попереднє зашумлення частини даних. Частина приватних записів (медичних, фінансових, тощо) піддається DP-обробці на внутрішній стороні – шляхом застосування механізмів Лапласа або Гауса до агрегованих статистик, побудови синтетичного датасета або проведення DP-тонкого налаштування допоміжної моделі;

– навчання тестових/захисних моделей зовнішніми підрядниками. Data scientist чи дослідник поза організацією отримує доступ лише до зашумлених даних або до публічного синтетичного датасета на їх основі. Він може експериментувати з архітектурами, відлагоджувати гіперпараметри та будувати початкові моделі, не маючи доступу до справжніх персональних даних;

– фінальне навчання ї моделі на чистих даних з обраним ε. Усередині організації модель, архітектура якої уже визначена на попередньому етапі, донавчається на справжніх даних із використанням DP-SGD. На цьому кроці обирається ε згідно з процедурою, описаною вище: за допомогою кривих «ε → (ризик, корисність)» та комбінованої функції втрат.

Як приклад можна розглянути сценарій із зображеннями, подібними до одного з випадків MNIST-датасету. Частина записів із чутливими мітками зашумлюється та доповнюється публічними зразками, на основі чого формується синтетичний датасет для внутрішніх або зовнішніх експериментів. Лише на останньому етапі, коли архітектура та гіперпараметри моделі вже зафіксовані, виконується ДК-навчання на справжніх даних з ретельно підібраним ε.

Такий конвеєр мінімізує потребу в розкритті реальних даних, зменшує площу атаки для зловмисників і водночас дозволяє зберегти високу якість моделей, оскільки більшість експериментів проводиться на зашумлених або синтетичних даних.

Підсумовуючи результати третього розділу, основний акцент доцільно змістити з окремих атак на умови коректного застосування автоматизованого вибору ε.

По-перше, для відносно простих моделей, таких як невеликих MLP на табличних даних з великим train/test-gap автоматизований підбір ε має обмежену додаткову цінність. У таких сценаріях уже сам факт значного перенавчання та високих значень AUC для MIA однозначно вказує на потребу або зменшити складність моделі, або додати регуляризацію, або задати «достатньо малий» ε. Байєсівський оптимізатор у цьому випадку переважно формалізує очевидний компроміс і радше виконує калібрувальну, ніж критично необхідну функцію.

По-друге, автоматизований вибір ε є значно більш виправданим для глибоких, багатокомпонентних архітектур, насамперед трансформерів і LLM. У таких моделях класичні ознаки перенавчання можуть бути слабо виражені, тоді як локальна меморизація окремих прикладів і каналів витоку, як у випадку секретних рядків для GPT-2, залишається істотною. Саме тут комбінована цільова функція дає змогу не призначати ε вручну, а системно обирати такий рівень шуму, який одночасно обмежує ризик і зберігає прийнятну якість моделі на цільових задачах.

По-третє, для LLM та інших високоризикових застосувань (медичні, фінансові, корпоративні тексти) автоматизований вибір ε доцільно розглядати як невід'ємну частину процесу безпечного розгортання моделі. У цих умовах ручна фіксація ε є надто грубою, натомість оцінка кривих «ε → (accuracy, AUC, показники витоку)» і подальша оптимізація з урахуванням ваг ризику та корисності дає обґрунтований, відтворюваний вибір параметрів приватності.

Нарешті, у більш широкому плані результати експериментів свідчать, що автоматизований підбір ε має сенс насамперед там, де структура моделі й характер даних приховують справжній трейдоф «приватність–корисність». Для простих базових моделей достатньо класичних прийомів контролю перенавчання та грубого налаштування ε, тоді як для трансформерів та великих глибоких мереж саме байєсівська модель ризику

й автоматизований ε-оптимізатор надають якісно новий інструмент керування приватністю.

У третьому розділі було показано, що моделі без спеціального захисту є суттєво вразливими до атак на приналежність до вибірки. За наявності помітного розриву між тренувальною та тестувальною вибірками loss-атаки і shadow-атаки досягають значень AUC істотно вищих за 0,5, а пари TPR і FPR підтверджують здатність зловмисника правильно ідентифікувати значну частку учасників навчання за мінімальної кількості хибних спрацьовувань. Для чутливих доменів це означає, що вже базовий доступ до значень функції втрат або до векторів ймовірностей виходу моделі створює відчутний практичний ризик.

Запровадження диференційно-приватного стохастичного градієнтного спуску з обрізанням градієнтів радикально змінює ситуацію. Емпіричні криві залежності між ε, тестовою точністю та AUC атак демонструють стале зниження успішності loss- і shadow-атак у міру посилення шуму, тоді як корисність моделі втрачається не завжди однаково. Для задач із добре регуляризованими моделями та великими вибірками виявлено діапазони ε, де точність знижується незначно, а ризик атак майже не відрізняється від випадкового вгадування. У більш складних сценаріях із високою кількістю класів і початково великим перенавчанням досягнення схожого рівня захисту потребує суттєвішого зниження якості прогнозування.

Окремий аналіз було присвячено агрегованим багатокомпонентним моделям, до яких належать глибокі згорткові та мовні архітектури. Для таких систем глобальний розрив у функції втрат між учасниками й неучасниками здебільшого малий, тому класичні loss- і shadow-атаки демонструють AUC, близькі до 0,5, навіть без дуже сильного зашумлення. Водночас експерименти з відтворенням секретних рядків показали, що поодинокі витоки можливі, отже мале значення AUC не гарантує повної безпеки, а лише описує середній рівень ризику.

На основі цих спостережень у розділі було запропоновано розглядати вибір параметра ε як задачу багатокритеріальної оптимізації з композитною функцією втрат, де нормований ризик атак поєднується з втратою корисності моделі через вагові коефіцієнти. Побудовані криві для різних експериментів показують, що оптимальне ε істотно залежить від предметної області та прийнятної політики безпеки: у високоризикових сферах пріоритет має мінімізація ризику, у менш чутливих допускається більший компроміс на користь якості. Загальний висновок полягає в тому, що поєднання градієнтного кліпінгу, помірного шуму та ретельно підібраної архітектури дає змогу суттєво знизити практичний ризик атак на приналежність до вибірки й водночас зберегти модель придатною до використання у більшості реалістичних сценаріїв.

ВИСНОВКИ

У кваліфікаційній роботі виконано повний цикл дослідження, запланований у вступі. Сформульовано й реалізовано підхід до автоматизованого вибору параметра ε для механізмів диференційної конфіденційності на основі поєднання теоретичної моделі ризику та емпіричних результатів атак на приналежність до вибірки. Запропоновано систему метрик для кількісного оцінювання приватності та корисності моделей, побудовано композитну функцію втрат, що узгоджує ці показники, та проведено серію експериментів для моделей різної складності. Аналіз отриманих кількісних і якісних результатів показав, що розроблений підхід дає змогу виявляти діапазони значень ε, у яких емпіричний ризик атак за AUC наближається до рівня випадкового вгадування, тоді як показники якості моделі залишаються достатніми для практичного використання.

Виконана розробка органічно вписується у контекст сучасних вітчизняних і світових досліджень у галузі диференційної конфіденційності, але водночас пропонує їх подальший розвиток. На відміну від більшості існуючих підходів, де параметр ε задається як фіксований зовнішній параметр, у роботі він розглядається як змінна, що оптимізується через явну побудову компромісу між ризиком атак і втратою корисності. Такий ризик орієнтований підхід узгоджується з актуальними міжнародними тенденціями, проте доповнює їх практичною процедурою ухвалення рішень, що спирається на результати реальних атак та емпіричні криві «ε – ризик» і «ε – корисність» для конкретних моделей.

У межах роботи отримано низку результатів, які мають елементи наукової новизни. Сформульовано інтегральну модель ризику атак на приналежність до вибірки, що поєднує внесок різних типів атак, зокрема loss та shadow, у єдиній нормованій шкалі. Запропоновано спосіб побудови композитної функції втрат, яка узагальнює залежності між параметром ε, емпіричним ризиком та втратою корисності й дає змогу автоматизувати

вибір є відповідно до заданих пріоритетів приватності та якості. Експериментально досліджено поведінку диференційно-приватного навчання для моделей різної архітектури, встановлено характерні діапазони параметрів, де ризик атак істотно зменшується, а корисність зберігається на прийнятному рівні. Отримані результати створюють підґрунтя для підготовки наукових публікацій та подальших досліджень, пов’язаних із розширенням моделі ризику на інші класи атак, уточненням метрик корисності та застосуванням підходу до розподілених і потокових систем обробки даних.

Матеріали кваліфікаційної роботи можуть бути використані як у практичній діяльності, так і в навчальному процесі університету. Запропоновані методики аналізу ризику, сценарії атак на приналежність до вибірки, результати експериментів та приклади налаштування диференційно-приватного навчання становлять базу для розроблення лабораторних робіт, практичних завдань і спеціалізованих курсів з диференційної конфіденційності, захисту інформації та безпеки систем штучного інтелекту. З прикладової точки зору отримані підходи можуть бути інтегровані у внутрішні протоколи налаштування моделей в організаціях, що працюють із чутливими даними, і слугувати інструментом обґрунтування обраних значень є перед регуляторами, замовниками та користувачами. Сукупність виконаних теоретичних і практичних кроків свідчить про повноту виконання поставленого завдання та перспективність подальшого розвитку досліджень у цьому напрямі.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Calibrating noise to sensitivity in private data analysis / C. Dwork et al. *Theory of cryptography*. Berlin, Heidelberg, 2006. P. 265–284. URL: https://doi.org/10.1007/11681878_14 (date of access: 11.11.2025).

2. Erlingsson Ú., Pihur V., Korolova A. Rappor. *CCS'14: 2014 ACM SIGSAC conference on computer and communications security*, Scottsdale Arizona USA. New York, NY, USA, 2014. URL: https://doi.org/10.1145/2660267.2660348 (date of access: 11.11.2025).

3. Apple. URL: https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf (date of access: 11.11.2025).

4. Collecting telemetry data privately. *arXiv.org*. URL: https://arxiv.org/abs/1712.01524 (date of access: 11.11.2025).

5. Uber. URL: https://www.uber.com/en-HU/blog/building-ubers-multi-cloud-secrets-management-platform/ (date of access: 11.11.2025).

6. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). URL: http://data.europa.eu/eli/reg/2016/679/oj (date of access: 12.11.2025).

7. Ukoh D.-F., Adetunji M. AI act: the EU regulation. SSRN electronic journal. 2025. URL: https://doi.org/10.2139/ssrn.4607388 (date of access: 12.11.2025).

8. A decade of metric differential privacy: advancements and applications. *arXiv.org*. URL: https://arxiv.org/html/2502.08970v1 (date of access: 12.11.2025).

9. The Algorithmic Foundations of Differential Privacy / C. Dwork, A. Roth. *Foundations and Trends® in Theoretical Computer Science*. 2014. Vol. 9,

Iss. 3–4. P. 211–407. URL: https://doi.org/10.1561/0400000042 (date of access: 12.11.2025).

10. Differential Privacy: A Survey of Results / C. Dwork. *Theory and Applications of Models of Computation*. 2008. P. 1–19. URL: https://doi.org/10.1007/978-3-540-79228-4_1 (date of access: 12.11.2025).

11. Guidelines for Measuring and Reporting Differential Privacy / J.P. Near et al. *NIST Special Publication* 800-226. 2025. URL: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-226.pdf (date of access: 15.11.2025).

12. Differential Privacy: An Economic Method for Choosing Epsilon / J. Hsu et al. *ACM Conference on Computer and Communications Security (CCS)*. 2014. URL: https://haeberlen.cis.upenn.edu/papers/epsilon-csf2014.pdf (date of access: 15.11.2025).

13. The Role of Differential Privacy in GDPR Compliance / R. Cummings et al. 2018. URL: https://rachelcummings.com/wp-content/uploads/2018/09/GDPR_DiffPrivacy.pdf (date of access: 15.11.2025).

14. Membership Inference Attacks Against Machine Learning Models / R. Shokri et al. *2017 IEEE Symposium on Security and Privacy (SP)*. San Jose, CA, USA, 2017. P. 3–18. URL: https://arxiv.org/abs/1610.05820 (date of access: 20.11.2025).

15. Bayesian Estimation of Differential Privacy / R. Zanella-Béguelin et al. *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*. PMLR, 2023. URL: https://arxiv.org/abs/2206.05199 (date of access: 20.12.2025).

16. Adaptive Differential Privacy Preserving Based on Multi-Objective Optimization in Deep Neural Networks / X. Zeng et al. *Concurrency and Computation: Practice and Experience*. 2021. URL: https://onlinelibrary.wiley.com/doi/10.1002/cpe.6367 (date of access: 20.11.2025)

17. Understanding the Prior and the Posterior Distributions / S. Ahmed. Medium. 2020. URL: https://sarowarahmed.medium.com/understanding-the-prior-and-the-posterior-distributions-0f36f8737ecc (date of access: 21.11.2025).

18. ML-Leaks: Model and Data Independent Membership Inference Attacks / A. Salem et al. *NDSS*, 2019. URL: https://www.ndss-symposium.org/wp-content/uploads/2019/02/ndss2019_03A-1_Salem_paper.pdf (date of access: 30.11.2025).

19. GradDiff: Gradient-based Membership Inference Attacks / X. Wang et al. *Information Sciences*, 2024. URL: https://www.sciencedirect.com/science/article/pii/S0020025523016547 (date of access: 30.11.2025).

20. Shadow-Free Membership Inference Attacks / Y. Zhang et al. *IJCAI*, 2024. URL: https://www.ijcai.org/proceedings/2024/0639.pdf (date of access: 30.11.2025).

21. White-box Membership Inference Attacks against Diffusion Models / Y. Ye et al. *PoPETs*, 2025. URL: https://petsymposium.org/popets/2025/popets-2025-0068.pdf (date of access: 30.11.2025).