НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «ОДЕСЬКА ЮРИДИЧНА АКАДЕМІЯ»

ФАКУЛЬТЕТ КІБЕРБЕЗПЕКИ ТА ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

28 листопада 2025 року

КІБЕРБЕЗПЕКА В СУЧАСНОМУ СВІТІ: АКТУАЛЬНІ ВИКЛИКИ

МАТЕРІАЛИ VI МІЖНАРОДНОЇ НАУКОВО-ПРАКТИЧНОЇ КОНФЕРЕНЦІЇ

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Національний університет «Одеська юридична академія»
Факультет кібербезпеки та інформаційних технологій
Кафедра кібербезпеки

КІБЕРБЕЗПЕКА В СУЧАСНОМУ СВІТІ: АКТУАЛЬНІ ВИКЛИКИ

МАТЕРІАЛИ VI МІЖНАРОДНОЇ НАУКОВО-ПРАКТИЧНОЇ КОНФЕРЕНЦІЇ

28 листопада 2025 року, м. Одеса

Одеса, 2025

MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE
National University «Odesa Law Academy»
Faculty of Cybersecurity and Information Technologies
Department of Cybersecurity

CYBERSECURITY IN TODAY'S WORLD: ACTUAL CHALLENGES

MATERIALS OF THE VI INTERNATIONAL SCIENTIFIC AND PRACTICAL CONFERENCE

November 28, 2025, Odesa

Odesa, 2025

УДК 004.056

Відповідальній редактор:

О. В. Дикий, декан факультету кібербезпеки та інформаційних технологій, кандидат юридичних наук, доцент

Матеріали видано в авторській редакції.
Повну відповідальність за достовірність та якість поданого матеріалу несуть учасники конференції, їхні наукові керівники, які рекомендували ці матеріали до друку.

Рекомендовано до друку
Вченою радою Національного університету
«Одеська юридична академія»
(протокол №3 від 29.11.2025 р.)

Матеріали Міжнародної науково-практичної конференції «Кібербезпека в сучасному світі: актуальні виклики» (м. Одеса, 28 листопада 2025 р.). Одеса, 2025. 287 с.

У конференції взяли участь студенти, аспіранти, молоді вчені, викладачі та науковці. Конференція проводиться на базі Національного університету «Одеська юридична академія».

UDC 004.056

Editor-in-chief:

O. V. Dykyi, Dean of the Faculty of Cybersecurity and Information Technologies,
Candidate of Law, Associate Professor

The materials were published in the author's edition.
Full responsibility for the reliability and quality of the submitted material
bears the participants of the conference, their scientific supervisors,
who recommended these materials for publication.

Recommended for printing
by the Academic Council of the National University
«Odesa Law Academy»
(Minutes No. 3 dated 11/29/2025)

Materials of the International Scientific and Practical Conference «Cybersecurity in the Modern World: Current Challenges» (Odessa, November 28, 2025). Odesa, 2025. 287 p.

The conference was attended by students, postgraduates, young scientists, teachers and researchers. The conference is held at the National University «Odesa Law Academy».

VI Міжнародна науково-практична конференція

«Кібербезпека в сучасному світі: актуальні виклики»

28 листопада 2025 року

ТЕМАТИЧНІ НАПРЯМКИ КОНФЕРЕНЦІЇ:

Секція 1. Алгоритмічні та технічні аспекти кібербезпеки

Секція 2. Штучний інтелект та інформаційні технології

Секція 3. Управлінські, соціальні та психологічні аспекти взаємодії з кіберпростором

Секція 4. Нормативно-правові засади кібербезпеки та захисту інформації

E-mail: kiberprostirconf@gmail.com (Для питань та тез)

VI International Scientific and Practical Conference

«Cybersecurity in today's world: actual challenges»

28 November 2025 year

THEMATIC DIRECTIONS OF THE CONFERENCE:

Section 1. Algorithmic and technical aspects of cybersecurity

Section 2. Artificial intelligence and information technologies

Section 3. Management, social and psychological aspects of interaction with cyberspace

Section 4. Regulatory and legal principles of cybersecurity and information protection

E-mail: kiberprostirconf@gmail.com (Для питань та тез)

АНАЛІЗ СУЧАСНИХ МЕХАНІЗМІВ ЗАХИСТУ ВІД АТАК НА ВИЯВЛЕННЯ НАЛЕЖНОСТІ

Хоменко Ігор

аспірант кафедри кібербезпеки Національного технічного університету «Харківський політехнічний інститут»

Широке впровадження машинного навчання (ML та великих мовних моделей (LLM), які навчаються на величезних наборах даних, що часто містять конфіденційну або приватну інформацію, додало нові вектори атак та загрози для приватності. Однією з фундаментальних загроз є атаки на визначення належності (Membership Inference Attack), під час яких зловмисник намагається встановити, чи був конкретний запис використаний у навчальному наборі моделей. Наприклад, якщо зловмисник визначить, що медичний запис пацієнта був використаних для навчання моделі прогнозування захворювань, це розкриє діагнози пацієнта. Успішна атака на визначення належності може мати критичні наслідки та завдати значні фінансові збитки [1].

Метою даного дослідження є проведення комплексного аналізу атак на визначення належності на сучасні системи штучного інтелекту, серед яких є великі мовні моделі, які впроваджуються у всі галузі нашого життя. Дослідження має на меті ідентифікувати та оцінити ефективність різноманітних векторів атак, а також проаналізувати відповідні механізми та методи захисту [2]. Під час дослідження було проведено систематичний огляд та аналіз ключових наукових праць, технічних звітів та виконано експеримент за допомогою llm-jp-membership-inference на ресурсі Hugging Face, що надає віртуальну машину і мовні моделі для експериментів [3].

Проведене дослідження дозволило виявити та кількісно оцінити ключові аспекти витоку інформації про належність у різних парадигмах машиного навчання. Зловмисник навчає модель атаки (класифікатор), використовуючи вектори передбачень (prediction vectors) або значення впевненості (confidence values) цільової моделі. Модель атаки розрізняє поведінку цільової моделі на вхідних даних, використаних для її навчання (in), та даних, які вона не бачила і прогнозує (out). Для навчання моделі атаки часто використовується метод тіньоового навчання (shadow training), де створюються допоміжні тіньові моделі зі схожою архітектурою та розподілом даних. Результати підтверджують, що успіх MIA прямо пов'язаний зі ступенем перенавчання цільової моделі. Якщо модель схильна до перенавчання (тобто демонструє високу точність на навчальному наборі, але низьку на тестовому), вона випускає більше інформації про свої навчальні вхідні дані [3].

Захисні стратегії проти атак на виявлення належності варіюються залежно від середовища навчання та цілей, які вони переслідують, часто вимагаючи компромісу між точністю та приватністю. Середовища можуть бути

типу Централізованого навчання (CL), федеративного навчання (FL) та доналаштування (Fine-Tuning LLM). Порівняльний аналіз захисних методів наведений у Таблиці 1 на основі тестування за допомогою репозиторію llm-jp-membership-inference та технічних звітів інших методів захисту [4].

Сутність методу Диференціальної приватності (DP) полягає у введенні ретельно каліброваного шуму в чутливі дані, градієнти або параметри моделі. Це є "легкою" технікою, що забезпечує строгий захист приватності, обмежуючи розбіжність між ймовірністю отримання результату, коли індивідуальний зразок включено до навчального набору, і коли його виключено. Завдяки такій конструкції, моделі, захищені DP, за визначенням є стійкими до атак на виявлення належності (MIA) [6].

Механізм регуляризації є однією з ключових стратегій захисту проти атак на виявлення належності (MIA), оскільки він безпосередньо спрямований на усунення основної причини витоку приватності — перенавчання (overfitting) моделі [7]. Регуляризація ефективно захищає від MIA, оскільки вона прагне зменшити рівень перенавчання моделей машинного навчання, що обмежує їхню прогностичну силу та узагальнювальну здатність. Моделі з гарною регуляризацією повинні мінімізувати витік інформації про свої навчальні дані, і атака MIA може використовуватися як метрика для кількісної оцінки цього рівня захисту. Регуляризаційні методи прагнуть зменшити розрив в узагальненні (generalization gap), тим самим знижуючи вразливість до MIA [8].

Обфускація вихідних даних (або пертурбація вихідних даних) є механізмом захисту, який застосовується на фазі висновку (inference) моделі. Основною метою цього підходу є зменшення кількості інформації, яку зловмисник може отримати з прогнозованого вектору ймовірностей (prediction vector) цільової моделі, щоб ускладнити проведення атаки MIA. Цей механізм є пост-процесинговим методом, який використовується для узгодження вихідних даних моделі між навчальними та тестовими зразками[9].

Селективна обфускація даних (SOFT) є інноваційним механізмом захисту, спеціально розробленим для мінімізації витоку приватності під час тонкого налаштування великих мовних моделей [10]. Суть методу полягає у використанні селективного вибору впливових даних для зменшення ризику атак на виявлення належності. SOFT реалізується шляхом заміни найбільш вразливих зразків у наборі даних для тонкого налаштування на обфусковані парафрази [11].

Основний принцип Безпечної агрегації (Secure Aggregation) полягає у запобіганні витоку інформації шляхом обміну зашифрованими оновленнями моделі замість відкритих [12]. SecAgg, зокрема, використовує методи розподілу секретів (secret sharing) та подвійного маскування. Це забезпечує, що центральний сервер не має доступу до індивідуальних локальних оновлень моделей клієнтів. Завдяки цьому, центральний агрегатор може обчислити

лише суму або середнє значення оновлень, отримуючи знання виключно про агреговану модель [13].

Часткове спільне використання є стратегією захисту у федеративному навчанні, головна мета якої — зменшити ефективність атак на виявлення належності (MIA) шляхом придушення або обмеження обсягу інформації, якою обмінюються під час комунікаційної фази. Цей підхід зменшує обсяг даних, доступних зловмисникам, обмежуючи обмін параметрами або градієнтами моделі [15].

Таблиця 1. Порівняльний аналіз захисних механізмів проти MIA

№	Механізм захисту	Контекст	Захист	Фаза Застосування
1	Диференційна приватність	CL/FL	Додавання шуму до даних, градієнтів або параметрів моделі. Забезпечує строгі теоретичні гарантії приватності	Навчання
2	Регуляризація (L2 або Dropout)	CL	Зменшення перенавчання (overfitting), яке є основною причиною витоку належності	Навчання
3	Обфускація вихідних даних (Output Perturbation)	CL	Округлення точності вектора прогнозування або обмеження вихідних даних моделі до k найбільш імовірних класів	Висновок
4	Дистиляція знань (Knowledge Distillation)	CL	Передача знань від більшої моделі-вчителя до меншої моделі-учениці, що сприяє кращому узагальненню та зниженню перенавчання	Навчання
5	Селективна обфускація даних	Fine-tuning LLM	Виявлення та заміна найбільш впливових/вразливих зразків навчального набору семантично еквівалентними парафразами	Навчання
6	Безпечна агрегація	FL	Використання криптографічних методів (наприклад, Secret Sharing) для забезпечення того, що центральний сервер бачить лише агрегований результат, а не індивідуальні оновлення клієнтів	Комунікація та агрегація
7	Часткове спільне використання	FL	Приховування певних оновлень моделі (наприклад, селективний градієнт) для зменшення ефективності атак	Комунікація

Аналіз поточних механізмів захисту від атак на виявлення належності (MIA) демонструє, що всі вони оперують у складному полі компромісів між

конфіденційністю (privacy), функціональною корисністю моделі (utility) та обчислювальною ефективністю. Жоден з існуючих методів не забезпечує універсального, бездоганного захисту без певних обмежень. Хоча DP є золотим стандартом, оскільки надає строгі теоретичні гарантії захисту від MIA, її впровадження неминуче призводить до суттєвої втрати корисності моделі. Для досягнення надійного захисту (мале є) часто доводиться жертвувати точністю прогнозів. У федеративному навчанні (ФН), якщо кількість учасників невелика, додавання шуму може перешкоджати конвергенції глобальної моделі. Селективна обфускація даних (SOFT): Цей підхід є одним із найбільш збалансованих для захисту тонконастроєних ВММ. SOFT демонструє здатність значно знижувати успішність MIA (до рівня, близького до випадкового вгадування), при цьому мінімізуючи втрату корисності моделі. Його перевага полягає у селективній роботі з найбільш вразливими (впливовими) зразками, зменшуючи перенавчання моделі на їхніх точних формулюваннях через парафразування.

Подальші дослідження повинні бути спрямовані на розробку уніфікованих бенчмарків та комплексних механізмів, які зможуть одночасно забезпечувати високу приватність, корисність та ефективність.

Список використаної літератури:

1. Shokri R., Stronati M., Song C., Shmatikov V. Membership inference attacks against machine learning models. 2017 IEEE Symposium on Security and Privacy (SP). 2017. P. 3–18.
2. He J., Chen M., Gu Y., Chen K. Membership Inference Attacks on In-Context Learning Recommendation. Proceedings of the 17th ACM Conference on Recommender Systems (RecSys ’25). 2025. P. 1–12.
3. GitHub - Ihor-Khomenko/llm-jp-membership-inference: This is the repository for llm jp membership inference attack. GitHub. URL: https://github.com/Ihor-Khomenko/llm-jp-membership-inference (date of access: 01.11.2025).
4. Puerto H., Gubri M., Yun S., Oh S. J. Scaling Up Membership Inference: When and How Attacks Succeed on Large Language Models. arXiv preprint arXiv:2411.00154. 2024.
5. Zhang K., Cheng S., Guo H. et al. SOFT: Selective Data Obfuscation for Protecting LLM Fine-tuning against Membership Inference Attacks. arXiv preprint. 2024.
6. Meeus M., Shilov I., Jain S. et al. SoK: Membership Inference Attacks on LLMs are Rushing Nowhere (and How to Fix It). arXiv preprint arXiv:2406.17975. 2024.
7. Nasr M., Shokri R., Houmansadr A. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. 2019 IEEE Symposium on Security and Privacy (SP). 2019. P. 739–753
8. Dwork C. Differential privacy. Encyclopedia of Cryptography and Security. Springer, 2011. P. 338–340.

9. Carlini N., Tramer F., Wallace E. et al. Extracting Training Data from Large Language Models. 30th USENIX Security Symposium (USENIX Security 21). 2021. P. 2633–2650.
10. Yeom S., Giacomelli I., Fredrikson M., Jha S. Privacy risk in machine learning: Analyzing the connection to overfitting. 2018 IEEE 31st Computer Security Foundations Symposium (CSF). 2018. P. 268–282.
11. Hu H., Salcic Z., Sun L. et al. Membership Inference Attacks on Machine Learning: A Survey. ACM Computing Surveys (CSUR). Vol. 54, No. 11s. 2022. P. 1–37.
12. Duan M., Suri A., Mireshghallah N. et al. Do membership inference attacks work on large language models? arXiv preprint arXiv:2402.07841. 2024.
13. Maini P., Jia H., Papernot N., Dziedzic A. LLM Dataset Inference: Did you train on my dataset? arXiv preprint arXiv:2406.06443. 2024
14. Bonawitz K., Ivanov V., Kreuter B. et al. Practical secure aggregation for privacy-preserving machine learning. Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. 2017. P. 1175–1191.
15. Yuan W., Yang C., Nguyen Q. V. H. et al. Interaction-level Membership Inference Attack Against Federated Recommender Systems. Proceedings of the Web Conference 2023. 2023. P. 1–10.

Ключові слова: Атака на виявлення належності, Великі мовні моделі, приватність, захисні механізми.
Keywords: Membership Inference Attack, Large Language Models, Privacy, Defence Mechanisms

ВИЯВЛЕННЯ НЕТИПОВИХ ПОДІЙ У СИСТЕМАХ ВІДЕОСПОСТЕРЕЖЕННЯ

Чикунов Павло

кандидат технічних наук, доцент кафедри кібербезпеки Національного університету «Одеська юридична академія», доцент

Флюнт Владислав

студент Національного університету «Одеська юридична академія»

Системи відеоспостереження є одним із інструментів моніторингу та забезпечення безпеки у громадських місцях, транспортних вузлах, промислових об’єктах, навчальних закладах і приватних структурах. Зі зростанням кількості камер та обсягу даних постає проблема оперативного аналізу відеопотоків. Оператор фізично не може відстежувати десятки джерел одночасно, тому автоматизовані системи відеоспостереження є критично необхідними.

Використання технологій штучного інтелекту дозволяє автоматизувати процеси виявлення нетипових або аномальних подій, таких як залишені предмети, поява людини у забороненій зоні, натовпова паніка, агресивна поведінка або порушення маршрутів руху. Ці події часто не піддаються формальному опису, що робить традиційні правило-орієнтовані методи