© Bondarovets S, Okhrimenko T. Modern artificial intelligence models protection methods from adversarial attacks // *Ukrainian Scientific Journal of Information Security, 2025, vol. 31, issue 3, pp. 142-150.* ©

DOI: 10.18372/2225-5036.31.21160

СУЧАСНІ МЕТОДИ ЗАХИСТУ МОДЕЛЕЙ ШТУЧНОГО ІНТЕЛЕКТУ ВІД ЦІЛЕСПРЯМОВАНИХ АТАК

Сергій Бондаровець, Тетяна Охріменко

Державний університет «Київський авіаційний інститут», Київ, Україна

БОНДАРОВЕЦЬ Сергій Сергійович
Рік та місце народження: 1995 рік, м. Біла Церква, Україна.
Освіта: Національний авіаційний університет, 2018 рік.
Посада: аспірант Державного університету «Київський авіаційний інститут».
Наукові інтереси: штучний інтелект, інформаційна безпека, кібербезпека.
Публікації: 2 наукові статті.
E-mail: 23933@stud.kai.edu.ua
Orcid: 0009-0007-2205-6768

ОХРІМЕНКО Тетяна Олександрівна, к.т.н., ст. дослідник
Рік та місце народження: 1990 рік, м. Вінниця, Україна.
Освіта: Національний авіаційний університет, 2012 рік.
Посада: заступник декана з наукової роботи Факультету комп'ютерних наук та технологій КАІ.
Наукові інтереси: інформаційна безпека, реагування на інциденти інформаційної безпеки, квантова криптографія.
Публікації: більше 100 наукових публікацій, серед яких монографії, наукові статті, матеріали та тези доповідей на конференціях, патенти та авторські свідоцтва.
E-mail: t.okhrimenko@npp.kai.edu.ua
ORCID: 0000-0001-9036-6556

Анотація. *У статті представлено комплексний огляд сучасного стану безпеки моделей штучного інтелекту (ШІ), систематизуючи вектори цілеспрямованих атак та відповідні методи захисту. Проаналізовано еволюцію ландшафту загроз, починаючи від класичних атак на моделі машинного навчання (ML) і закінчуючи специфічними вразливостями, притаманними сучасним великим мовним моделям (LLM). Вступна частина окреслює актуальність проблеми в контексті глибокої інтеграції ШІ в критичні інфраструктури та бізнес-процеси, підкреслюючи перехід від реактивного виправлення вразливостей до проактивного управління ризиками, що відображено в галузевих стандартах, таких як NIST AI Risk Management Framework. Основна частина дослідження починається з детальної класифікації модально-агностичних атак, включаючи змагальні атаки (adversarial examples), отруєння даних (data poisoning), впровадження бекдорів (backdoors), викрадення моделей (model stealing), атаки на визначення належності (membership inference) та інверсію моделі (model inversion). Далі проводиться аналіз найбільш критичних загроз для сучасних систем за критеріями поширеності, потенційної шкоди та складності виявлення, акцентуючи увагу на атаках на ланцюг постачання та витоках даних. Систематизовано методи захисту загального призначення, структуровані за етапами життєвого циклу моделі: на рівні даних (санітизація, диференційна приватність), під час навчання (змагальне навчання, робастна оптимізація) та на етапі експлуатації (моніторинг, політики безпеки). Окремий розділ присвячено парадигмальному зсуву, спричиненому LLM. Детально розглядаються LLM-специфічні загрози: ін'єкції запитів (prompt injection), включаючи прямі (jailbreaks) та непрямі атаки в RAG-системах, маніпуляції з виводом, а також ризики, пов'язані з донавчанням та використанням зовнішніх інструментів. Відповідно, аналізуються багаторівневі стратегії захисту для LLM, такі як посилення системних підказок, впровадження захисних бар'єрів (guardrails), red-teaming, безпечне проєктування RAG-систем та інструментів. Аналітичний синтез узагальнює сильні та слабкі сторони розглянутих підходів, оцінює ризики хибних спрацьовувань та пропусків (FP/FN) і умови їх застосування. У висновках підсумовано ключові результати та визначено перспективні напрями для подальших досліджень, зокрема розробку формальних гарантій безпеки для LLM, стандартизацію бенчмарків та розвиток комплексних систем моніторингу.*
Ключові слова: *штучний інтелект; кібербезпека; захист штучного інтелекту; великі мовні моделі; цілеспрямовані атаки.*

Вступ

Стрімка інтеграція систем штучного інтелекту (ШІ) в критично важливі сфери, від автономного транспорту та медичної діагностики до фінансових ринків та оборонних технологій, перетворила питання їхньої безпеки з академічної проблеми на нагальну суспільну та економічну необхідність [1]. Здатність моделей машинного навчання (ML) генерувати високоякісний контент, аналізувати складні дані та автоматизувати прийняття рішень супроводжується появою нових, унікальних векторів атак. Ці атаки можуть призводити не лише до фінансових збитків чи репутаційної шкоди, але й до фізичної небезпеки та підриву довіри до технологій загалом.

Актуальність проблеми безпеки ШІ підкреслюється активною розробкою та впровадженням галузевих стандартів і фреймворків. Такі ініціативи, як Artificial Intelligence Risk Management Framework (AI RMF 1.0) від Національного інституту стандартів і технологій США (NIST), знаменують собою фундаментальний зсув у підходах до безпеки [3]. Вони переводять фокус із суто технічних аспектів вразливостей на комплексну, соціально-технічну систему управління ризиками, що охоплює весь життєвий цикл ШІ-

системи – від проєктування до виведення з експлуатації [3].

AI RMF визначає управління ризиками як ключовий компонент відповідальної розробки та використання ШІ, що сприяє підвищенню надійності (trustworthiness) систем [3].

Поява та широке розповсюдження генеративних моделей, зокрема великих мовних моделей (LLM), ще більше розширило поверхню атаки, породивши нові класи загроз, нехарактерні для традиційних ML-систем [2]. У відповідь на це, спільноти, такі як Open Web Application Security Project (OWASP), розробили спеціалізовані переліки ризиків, наприклад, OWASP Top 10 for Large Language Model Applications, що систематизують такі загрози, як ін'єкції запитів (prompt injection) та надмірне надання повноважень (excessive agency) [4].

Ця еволюція від дослідження ізольованих атак, як-от оманливі приклади (adversarial examples) для класифікаторів зображень, до створення комплексних фреймворків управління ризиками свідчить про зрілість галузі безпеки ШІ. Сучасний підхід вимагає не просто розробки точкових захисних механізмів, а побудови цілісних, ешелонованих систем безпеки, що враховують як технічні, так і організаційні аспекти.

Об'єктом даного огляду є процеси функціонування моделей ШІ, що піддаються цілеспрямованим атакам. Предметом дослідження виступають сучасні методи та стратегії захисту моделей ШІ на різних етапах їхнього життєвого циклу.

Внесок цього огляду полягає у систематизації та критичному аналізі актуального стану досліджень у сфері безпеки ШІ. Він поєднує фундаментальні знання про атаки на традиційні моделі ML із глибоким аналізом нових загроз, специфічних для LLM-систем. Стаття надає структурований огляд захисних механізмів, оцінюючи їхні переваги, недоліки та доцільність застосування в різних контекстах, що робить її корисною для дослідників, інженерів та фахівців з кібербезпеки, які працюють над створенням надійних та безпечних систем ШІ.

Постановка проблеми

Широке впровадження систем ШІ у практичні завдання, від обробки природної мови до керування автономними системами, створює нові виклики для інформаційної безпеки. На відміну від традиційних програмних систем, де вразливості часто пов'язані з помилками в коді, системи ШІ демонструють принципово нові поверхні атаки, що випливають із самої природи їхнього навчання на даних. Проблема полягає в тому, що зловмисники можуть цілеспрямовано маніпулювати як навчальними даними, так і вхідними запитами до вже навченої моделі, щоб викликати непередбачувану, помилкову або шкідливу поведінку. Це створює прямий зв'язок із важливими практичними завданнями забезпечення надійності, конфіденційності та цілісності критичних систем, що вимагає систематичного дослідження та розробки спеціалізованих методів захисту.

Аналіз останніх досліджень і публікацій.

Наукова література останніх років демонструє значний прогрес у розумінні та протидії загрозам безпеці ШІ. Ранні оглядові роботи переважно фокусувалися на цілеспрямованих атаках (adversarial attacks) у сфері комп'ютерного зору, детально класифікуючи методи генерації оманливих прикладів та перші спроби захисту, такі як змагальне навчання [6]. Ці дослідження заклали фундаментальну основу, визначивши ключові поняття, такі як моделі загроз (білий, сірий, чорний ящик) та цілі атак (цільові та нецільові) [8].

З часом фокус досліджень розширився, охоплюючи інші класи атак. З'явилися комплексні огляди, присвячені атакам на приватність, зокрема атакам на визначення належності (membership inference attacks, MIA) та викраденню моделей (model stealing) [9]. Дослідження MIA виявили, що моделі можуть «запам'ятовувати» та опосередковано розкривати чутливу інформацію з навчальних даних, що становить серйозну загрозу приватності [11]. Водночас аналіз атак з викрадення моделей продемонстрував економічні ризики, пов'язані з втратою інтелектуальної власності, коли зловмисник може відтворити функціональність пропрієтарної моделі, маючи до неї лише API-доступ [10].

Важливим напрямом стало вивчення атак на цілісність даних та моделей, зокрема отруєння даних (data poisoning) та впровадження бекдорів (backdoor attacks) [14]. Огляди в цій галузі показали, як зловмисники можуть маніпулювати навчальними даними для створення прихованих вразливостей, які активуються за допомогою спеціальних тригерів на етапі експлуатації [4]. Це підкреслило критичну важливість безпеки всього ланцюга постачання ШІ (AI supply chain security).

Паралельно з каталогізацією атак, значна увага приділялася розробці та аналізу захисних механізмів. Огляди, присвячені захисту, систематизували такі підходи, як диференційна приватність (differential privacy), що забезпечує формальні гарантії приватності [16], та сертифікована стійкість (certified robustness), що надає доказові гарантії стійкості до цілеспрямованих атак у певних межах [18]. Однак ці ж дослідження виявили фундаментальний компроміс між надійністю гарантій, корисністю моделі (utility) та обчислювальною складністю, що обмежує практичне застосування багатьох формальних методів [19].

Поява LLM спричинила новий виток досліджень, що знайшло відображення у створенні таких фреймворків, як OWASP Top 10 for LLM Applications [5]. Ці документи стали першою спробою систематизувати нові, унікальні для LLM вектори атак, як-от ін'єкції запитів. Проте, як зазначають останні огляди, існує помітний розрив між швидкістю розвитку LLM-систем (особливо агентних) та формалізацією методів їх захисту [11]. Досі бракує систематичних досліджень, які б комплексно аналізували безпеку складних RAG-систем (Retrieval-Augmented Generation) та ланцюгів інструментів (tool chains), що є раніше невирішеною частиною загальної проблеми [21]. Сучасна наукова

література перебуває на етапі, коли вона намагається наздогнати технологічний прогрес, адаптуючи відомі принципи безпеки до нової парадигми генеративного ШІ та ідентифікуючи принципово нові виклики.

Мета даного дослідження полягає у систематизації та критичному аналізі сучасних методів захисту моделей штучного інтелекту від цілеспрямованих атак, з акцентом на еволюцію загроз та захисних стратегій у контексті переходу до великомасштабних генеративних моделей.

Основна частина дослідження

Загальна класифікація атак на моделі штучного інтелекту

Розуміння ландшафту загроз є першим кроком до побудови ефективної системи захисту. Атаки на моделі ШІ можна класифікувати за їхньою метою, необхідним рівнем знань про модель та етапом життєвого циклу, на якому вони здійснюються. Нижче наведено таксономію основних, модально-агностичних атак, що застосовуються до широкого спектра ML-систем.

1. **Оманливі приклади / Атаки ухилення (Adversarial Examples / Evasion Attacks)**: Це вхідні дані для моделей машинного навчання, які зловмисник навмисно розробив, щоб змусити модель зробити помилку; вони схожі на оптичні ілюзії для машин [6]. Ці атаки відбуваються на етапі інференсу (експлуатації). Їхня мета – змусити модель зробити помилкове передбачення шляхом внесення вхідних даних невеликих, часто непомітних для людини, збурень [7]. Класичним прикладом є незначна модифікація пікселів зображення, що змушує класифікатор розпізнати один об'єкт як інший. Атаки ухилення поділяються за кількома критеріями:
    a. за рівнем доступу до моделі: white-box (зловмисник має повний доступ до архітектури та параметрів моделі), black-box (доступ лише до API моделі для отримання передбачень) та grey-box (часткові знання про модель) [8].
    b. за метою: untargeted (мета – будь-яка помилкова класифікація) та targeted (мета – змусити модель видати конкретний, заздалегідь визначений неправильний результат) [8].
2. **Отруєння даних (Data Poisoning):** цей тип атак спрямований на етап навчання моделі. Зловмисник впроваджує шкідливі (отруєні) дані в навчальний набір, щоб погіршити загальну продуктивність моделі або, що небезпечніше, створити приховані вразливості [4]. Наприклад, зловмисник може додати в набір даних для розпізнавання спаму листи зі шкідливими посиланнями, навмисно помічені як «не спам», щоб навчити модель ігнорувати подібні загрози в майбутньому.
3. **Бекдори / Троянські атаки (Backdoors / Trojan Attacks)**: це витончена форма отруєння даних. Мета зловмисника – не просто погіршити модель, а вбудувати в неї приховану, шкідливу поведінку, яка активується лише за наявності специфічного тригера (trigger) у вхідних даних [14]. В інший час модель поводиться абсолютно нормально, що робить атаку надзвичайно складною для виявлення. Наприклад, модель розпізнавання дорожніх знаків може бути навчена ігнорувати знак «Стоп», якщо в кутку зображення присутній невеликий жовтий квадрат (тригер) [1].
4. **Викрадення / Екстракція моделі (Model Stealing / Extraction):** атака, спрямована на порушення конфіденційності та крадіжку інтелектуальної власності. Зловмисник, маючи лише black-box доступ до API моделі, систематично надсилає запити та аналізує відповіді, щоб навчити власну, «клоновану» модель, яка функціонально еквівалентна оригінальній [10]. Це дозволяє не лише безкоштовно користуватися комерційною моделлю, але й використовувати її для розробки ефективніших цілеспрямованих атак [12].
5. **Атаки на визначення належності (Membership Inference Attacks, MIA)**: це атака на приватність, метою якої є визначення, чи був конкретний запис даних використаний під час навчання цільової моделі [9]. Успішна атака може розкрити чутливу інформацію, наприклад, чи були медичні дані певної особи використані для навчання моделі діагностики захворювань, що є прямим порушенням конфіденційності [9].
6. **Інверсія моделі (Model Inversion):** ще одна серйозна атака на приватність, яка йде далі за MIA. Її мета – не просто визначити належність даних, а реконструювати частини оригінальних навчальних даних або їхні характерні риси, маючи доступ до моделі [24]. Наприклад, з моделі розпізнавання облич можна спробувати відновити усереднені зображення облич людей з навчального набору, що відповідають певним класам [24].
7. **Витік даних / Порушення приватності (Privacy / Data Leakage):** це загальна категорія загроз, коли модель ненавмисно розкриває чутливу інформацію зі своїх навчальних даних у своїх відповідях. Це може статися через «запам'ятовування» (memorization) моделлю унікальних або рідкісних фрагментів даних [5].
8. **Контамінація даних / Атаки на ланцюг постачання (Data Contamination / Supply Chain Attacks):** це широка категорія загроз, що охоплює будь-яке компрометування компонентів, які використовуються для створення або функціонування ШІ-системи. Це може бути використання скомпрометованих сторонніх бібліотек, наборів даних або, що особливо актуально сьогодні, попередньо навчених моделей з неперевірених джерел [4].
9. **Зсув розподілу / Дані поза розподілом (Distribution Shift / Out-of-Distribution, OOD):** Це не завжди зловмисна атака, а скоріше фундаментальна проблема надійності. Модель, навчена на даних з одного розподілу, може значно втратити в продуктивності, зіткнувшись на етапі експлуатації з даними, що мають інший статистичний розподіл. Зловмисники можуть свідомо експлуатувати цю вразливість, подаючи моделі OOD-дані, щоб

викликати непередбачувану або помилкову поведінку [26].

Такий підхід до класифікації дозволяє не лише зрозуміти природу кожної загрози, але й визначити, на якому етапі життєвого циклу ШІ-системи необхідно впроваджувати відповідні захисні механізми.

Критичні вектори атак на сучасні системи

Хоча всі перелічені атаки становлять загрозу, деякі з них є особливо небезпечними для сучасних, широко розгорнутих систем ШІ через поєднання кількох факторів: високий потенціал шкоди, складність виявлення, тривалий час відновлення та здатність слугувати плацдармом для подальших атак. Аналіз показує, що найбільш критичними є атаки, які підривають фундаментальну довіру до моделі та її ланцюга постачання.

Отруєння даних та атаки на ланцюг постачання виділяються як одна з найсерйозніших загроз. На відміну від атак ухилення, які є тимчасовими та діють на окремі запити, отруєння є персистентною компрометацією самої моделі [15]. Одного разу отруєна модель стає, по суті, шкідливим активом, який може поширюватися і використовуватися в незліченній кількості downstream-застосунків. Виявлення таких атак є надзвичайно складним, оскільки модель може демонструвати високу точність на стандартних тестах, а її шкідлива поведінка (наприклад, бекдор) залишається прихованою до активації тригером [14]. Відновлення вимагає повної перевірки всіх навчальних даних та повного перенавчання моделі, що є надзвичайно дорогим і тривалим процесом, особливо для великомасштабних моделей. Ця загроза посилюється сучасною практикою використання попередньо навчених моделей з публічних репозиторіїв, що створює значні ризики для ланцюга постачання [4].

Викрадення моделі є критичною загрозою з двох причин. По-перше, це пряма економічна шкода через втрату інтелектуальної власності, яка може коштувати мільйони доларів інвестицій у дані та обчислювальні ресурси [10]. По-друге, що є ще більш небезпечним, викрадення моделі є потужним інструментом розвідки для зловмисника [22]. Маючи точну копію цільової моделі, атакуючий може проводити необмежену кількість white-box атак офлайн для розробки ідеальних оманливих прикладів або пошуку вразливостей приватності. Це значно підвищує ефективність подальших атак ухилення чи інверсії моделі, роблячи їх майже невідворотними.

Витоки даних та атаки на приватність (MIA та інверсія моделі) становлять екзистенційну загрозу для багатьох застосувань ШІ, особливо в таких регульованих галузях, як охорона здоров'я та фінанси. Фундаментальна властивість глибоких нейронних мереж «запам'ятовувати» унікальні риси навчальних даних створює постійний ризик їх ненавмисного розкриття [11]. Наслідки таких витоків виходять за межі технічної площини, призводячи до серйозних юридичних санкцій (наприклад, згідно з GDPR), фінансових штрафів та, що найважливіше, до повної втрати довіри користувачів [17]. Складність виявлення та неможливість «видалити» знання з уже навченої моделі без складних процедур (machine unlearning) роблять ці атаки особливо критичними.

Таким чином, відбувається зміна парадигми оцінки критичності загроз. Найнебезпечнішими є не ті атаки, що викликають миттєвий і очевидний збій (як проста атака ухилення), а ті, що діють приховано, підривають цілісність моделі на фундаментальному рівні та створюють умови для подальшої експлуатації. Це переводить фокус безпеки з захисту від окремих запитів на забезпечення надійності всього життєвого циклу ШІ-системи, від даних до розгортання.

Методи захисту загального призначення

Для протидії описаним загрозам розроблено низку захисних стратегій, які можна застосовувати на різних етапах життєвого циклу моделі ШІ. Ефективний захист вимагає багаторівневого підходу (defense-in-depth), що поєднує заходи на рівні даних, під час навчання та на етапі експлуатації.

Захист на рівні даних – це перший і найважливіший рубіж оборони, спрямований на запобігання атакам отруєння та контамінації.

– Санітизація та фільтрація даних (Data Sanitization and Filtering): цей підхід передбачає ретельну перевірку та очищення навчальних даних перед їх використанням. Методи включають виявлення аномалій та викидів (outlier detection) для ідентифікації підозрілих зразків, перевірку легітимності джерел даних та верифікацію цілісності даних протягом усього конвеєра обробки [4].

– Диференційна приватність (Differential Privacy, DP): це формальний математичний підхід до забезпечення приватності, який гарантує, що результат обчислення (наприклад, навчена модель) майже не зміниться, якщо з набору даних видалити або додати один будь-який запис. Найпоширенішим методом для глибокого навчання є DP-SGD (Differentially Private Stochastic Gradient Descent) [16].

Незважаючи на сильні теоретичні гарантії, DP має суттєві недоліки. По-перше, це компроміс «приватність-корисність»: вищий рівень приватності (більше шуму) зазвичай призводить до зниження точності моделі [17]. По-друге, дослідження показали, що цей спад точності є нерівномірним і непропорційно сильно впливає на недостатньо представлені (underrepresented) групи в даних, що може посилювати існуючу упередженість моделі [15].

Захист під час навчання – ці методи спрямовані на створення моделей, які є за своєю суттю більш стійкими до атак.

- Змагальне навчання (Adversarial Training): це один з найефективніших емпіричних методів захисту від атак ухилення. Ідея полягає в тому, щоб «імунізувати» модель, додаючи до навчального набору згенеровані оманливі приклади та навчаючи модель правильно їх класифікувати [6]. Це змушує модель вивчати більш робастні (стійкі) ознаки та згладжує її поверхню рішень.
- Робастна оптимізація (Robust Optimization): більш загальний клас методів, що модифікують процес оптимізації (навчання)

для мінімізації втрат у «найгіршому випадку» в межах певної області навколо кожної точки даних. Змагальне навчання є одним із прикладів робастної оптимізації.

- Сертифіковані підходи (Certified Defenses): на відміну від емпіричних методів, сертифіковані підходи надають математичні гарантії, що модель буде стійкою до будь-яких збурень у межах певного радіуса [18]. Ці методи, такі як interval bound propagation або randomized smoothing, є дуже потужними в теорії. Однак на практиці вони стикаються з серйозними проблемами:
  - масштабованість: багато сертифікованих методів є обчислювально дорогими і погано масштабуються на великі моделі, як-от LLM [14].
  - розрив з практикою: сертифікований радіус стійкості часто виявляється значно меншим, ніж практична стійкість, досягнута за допомогою емпіричних методів, як-от змагальне навчання [19].
  - вплив на точність: досягнення високого рівня сертифікованої стійкості часто призводить до значного падіння стандартної точності моделі [20].
- Моніторинг та детекція аномалій: системи моніторингу аналізують вхідні запити та вихідні дані моделі на предмет аномалій, які можуть свідчити про атаку. Це може включати виявлення OOD-даних, статистично нетипових запитів або спроб зловживання API [26].
- Обмеження частоти запитів (Rate Limiting): простий, але ефективний засіб проти атак викрадення моделі та атак типу «відмова в обслуговуванні» (Denial of Service). Обмеження кількості запитів з однієї IP-адреси або для одного користувача за певний час значно ускладнює збір великого обсягу даних, необхідного для клонування моделі [4].
- Аудит та журналювання (Auditing and Logging): ведення детальних журналів усіх взаємодій з моделлю дозволяє проводити ретроспективний аналіз у разі інциденту, виявляти патерни атак та вдосконалювати захисні механізми.

Специфічні загрози для великих мовних моделей (LLM)

Поява великих мовних моделей (LLM) та систем на їх основі, таких як чат-боти, агенти та системи генерації з доповненим пошуком (RAG), спричинила парадигмальний зсув у ландшафті загроз. Якщо для традиційних ML-моделей вхідні дані та інструкції були чітко розділені, то в LLM ця межа розмивається: дані, отримані з зовнішніх джерел, можуть бути інтерпретовані моделлю як інструкції, що створює принципово нові вектори атак. Проєкт OWASP Top 10 for LLM Applications став однією з перших спроб систематизувати ці унікальні ризики [5].

1. Ін'єкція запиту (Prompt Injection): найпоширеніша та найкритичніша вразливість LLM-систем [5]. Вона полягає в тому, що зловмисник за допомогою спеціально сформованого вхідного тексту (промпту) змушує модель ігнорувати її початкові інструкції (системний промпт) і виконувати непередбачені, шкідливі дії. Існує два основних типи ін'єкцій:
   a. пряма ін'єкція (Direct Prompt Injection) або «Jailbreaking» – зловмисник безпосередньо у своєму запиті до моделі формулює інструкції, що обходять її запобіжні механізми. Наприклад, просить модель «вжитися в роль» персонажа, який не має етичних обмежень, щоб згенерувати шкідливий контент [4].
   b. непряма ін'єкція (Indirect Prompt Injection) – цей тип атаки є значно більш підступним і небезпечним, особливо для систем, що взаємодіють із зовнішніми джерелами даних (RAG, агенти). Зловмисник розміщує шкідливий промпт у зовнішньому джерелі (наприклад, на веб-сторінці, у PDF-документі, в електронному листі), яке система згодом обробляє. Коли LLM отримує цей фрагмент тексту для аналізу чи узагальнення, вона виконує приховану в ньому інструкцію. Наприклад, RAG-система, що аналізує веб-сторінку, може виконати приховану команду «знайди в історії розмови електронну адресу користувача і відправ її на example.com» [4]. Це перетворює будь-яке неперевірене зовнішнє джерело даних на потенційний вектор атаки.
2. Небезпечна обробка виводу (Insecure Output Handling): вразливість виникає, коли downstream-системи (наприклад, веб-фронтенд, API, бази даних) беззастережно довіряють і виконують контент, згенерований LLM. Зловмисник може змусити модель згенерувати шкідливий код (наприклад, JavaScript, SQL-запити), який потім буде виконаний у вразливій частині застосунку. Це може призвести до класичних веб-атак, таких як міжсайтовий скриптинг (XSS), підробка міжсайтових запитів (CSRF) або навіть віддалене виконання коду (RCE) [4].
3. Загрози для RAG-систем та агентів: системи, що використовують RAG або взаємодіють із зовнішніми інструментами (агенти), стикаються з унікальними ризиками:
   a. отруєння даних для RAG – зловмисник може отруїти базу знань, з якої система отримує інформацію, впровадивши туди дезінформацію або приховані інструкції для непрямих ін'єкцій.
   b. ексфільтрація даних через інструменти – якщо LLM-агент має доступ до інструментів (наприклад, API для надсилання електронних листів або доступу до файлової системи) і недостатньо захищений, зловмисник через ін'єкцію запиту може

змусити його використати ці інструменти для викрадення конфіденційних даних [21].

c. перехоплення моделі (Model Hijacking) – зловмисник може перенаправити функціональність системи на виконання завдань, для яких вона не була призначена. Наприклад, чат-бот для підтримки клієнтів може бути використаний для написання фішингових листів або генерації шкідливого коду [21].

4. Контамінація під час донавчання (Fine-tuning contamination): специфічний для LLM варіант отруєння даних. Під час процесу донавчання (fine-tuning) або інструктивного донавчання (instruction tuning) для адаптації моделі до конкретних завдань, зловмисник може впровадити в набір даних невелику кількість шкідливих прикладів. Це може створити тонкі бекдори, прищепити моделі небажані упередження або змусити її генерувати дезінформацію за певних умов [14].

5. Ризики ланцюга постачання (Supply-chain risks): як і традиційні ML-системи, LLM-застосунки сильно залежать від сторонніх компонентів. Ризики включають використання скомпрометованих попередньо навчених базових моделей, вразливих плагінів або неперевірених наборів даних для донавчання. Вразливість в одному з цих компонентів може скомпрометувати всю систему [4].

Фундаментальна відмінність загроз для LLM полягає в тому, що атака експлуатує саму природу мови та контексту. Зловмисник маніпулює не математичними вразливостями моделі, а її здатністю до інтерпретації та виконання інструкцій, що робить традиційні методи захисту, орієнтовані на числові збурення, недостатньо ефективними. Безпека LLM-систем вимагає нового підходу, зосередженого на контролі інформаційних потоків, валідації даних на всіх етапах та управлінні повноваженнями моделі.

Багаторівневий захист для LLM-систем

Захист LLM-систем від специфічних загроз вимагає комплексного, багаторівневого підходу, який поєднує технічні засоби контролю, найкращі практики розробки та організаційні заходи. На відміну від захисту традиційних моделей, тут акцент зміщується з математичної стійкості на архітектурну безпеку та контроль взаємодії моделі з оточенням.

1. Посилення запитів та гігієна системних підказок (prompt hardening and hygiene): перший рівень захисту, що реалізується на етапі проєктування взаємодії з моделлю. Він включає розробку стійких до ін'єкцій системних підказок. Ефективні техніки включають:

a. чітке розмежування інструкцій та даних користувача, наприклад, за допомогою спеціальних роздільників або XML-тегів.

b. надання моделі явних інструкцій щодо того, як поводитися з потенційно шкідливими запитами.

c. використання технік, як-от «сендвічинг», коли дані користувача розміщуються між двома блоками надійних системних інструкцій.

Однак, слід пам'ятати, що посилення промптів є евристичним методом і не дає стовідсоткової гарантії захисту від цілеспрямованих атак [15].

2. Захисні бар'єри та політики як код (guardrails and policy-as-code): один з найважливіших архітектурних компонентів безпеки LLM. Guardrails – це незалежні модулі або моделі, які перевіряють як вхідні дані (запити користувача), так і вихідні дані (відповіді LLM) на відповідність заздалегідь визначеним політикам безпеки. Цей підхід дозволяє відокремити логіку безпеки від основної моделі, що робить систему більш надійною.

3. Red-Teaming та фреймворки оцінювання: Це проактивний підхід до виявлення вразливостей. Спеціалізовані команди (red teams) систематично намагаються «зламати» систему, використовуючи відомі та нові техніки атак, щоб виявити слабкі місця до того, як їх знайдуть зловмисники. Використання стандартизованих фреймворків для оцінки безпеки дозволяє кількісно виміряти стійкість системи до різних типів атак.

4. Безпечне використання інструментів та посилення RAG (RAG hardening): Для систем, що використовують зовнішні дані та інструменти, необхідні додаткові заходи:

a. фільтрація та санітизація отриманих даних – усі дані, отримані з зовнішніх джерел (вебсторінки, документи), повинні проходити через фільтри безпеки перед тим, як потрапити до LLM, для видалення потенційних непрямих ін'єкцій [21].

b. верифікація джерел та заземлення (grounding) – система повинна надавати перевагу перевіреним та надійним джерелам інформації. Відповіді моделі слід «заземлювати», тобто перевіряти, чи вони відповідають інформації з отриманих джерел, щоб зменшити ризик галюцинацій та дезінформації.

c. ізоляція та контроль доступу для інструментів – інструменти, до яких має доступ LLM, повинні працювати в ізольованому середовищі (sandbox) з мінімально необхідними привілеями (least privilege), щоб обмежити потенційну шкоду від успішної атаки [25].

5. Управління секретами та запобігання витоку даних (DLP): система не повинна зберігати або передавати у відкритому вигляді секрети (ключі API, паролі) в промптах. Необхідно використовувати спеціалізовані сховища секретів. Також слід впроваджувати DLP-сканери для моніторингу вихідного трафіку на предмет витоку конфіденційної інформації.

6. Безпека API та контроль доступу: кінцева точка API, через яку відбувається взаємодія з LLM, повинна бути захищена стандартними заходами веб-безпеки: надійна автентифікація, авторизація, шифрування трафіку (TLS), а також обмеження частоти запитів (rate limiting) та контроль сесій для запобігання зловживанням та DoS-атакам [4].

7. Водяні знаки та цифрові відбитки (watermarking and fingerprinting): хоча дані методи не є

прямим захистом від атак, вони можуть бути корисними для розслідування інцидентів. Вони дозволяють вбудовувати в згенерований контент непомітні маркери, які допомагають ідентифікувати, яка модель його створила, що може бути корисним для відстеження зловживань або викрадення моделі [28].

Ефективний захист LLM-системи – це не один інструмент, а комплексна стратегія, що поєднує заходи на рівні архітектури, даних, моделі та інфраструктури.

Переваги та недоліки сучасних методів захисту ШІ-систем

Аналіз сучасних методів захисту ШІ-систем виявляє фундаментальні компроміси між теоретичною надійністю, практичною застосовністю, продуктивністю та впливом на функціональність. Не існує універсального рішення; вибір оптимальної стратегії залежить від архітектури системи, моделі загроз та прийнятного рівня ризику.

Таблиця 1

Порівняльний аналіз підходів до захисту ШІ

Критерій	Формальні методи	Евристичні/Архітектурні підходи
Сильні сторони	Надають найвищий рівень математично доказових гарантій захисту від певних класів атак, що є незамінним для систем з високими вимогами до безпеки та приватності.	Є значно більш практичними, гнучкими та добре інтегруються в існуючі системи. Демонструють високу ефективність проти відомих атак.
Слабкі сторони	Непрактичність: обчислювально дорогі, погано масштабуються на LLM, майже завжди призводять до значного зниження точності моделі. Захищають лише від вузько визначених загроз.	Не надають формальних гарантій і можуть бути обійдені новими, невідомими техніками атак. Надійність залежить від якості реалізації та постійного оновлення.
Ризики False Positive/False Negative	(Не застосовується безпосередньо до методів навчання)	Високі ризики хибних спрацьовувань (FP), що шкодять користувацькому досвіду, та пропусків (FN), що призводять до інцидентів безпеки. Потребують ретельного балансування.
Сумісність	Обмежена для великомасштабних архітектур (LLM). Краще підходять для традиційних ML-моделей.	Є основним підходом для захисту LLM та RAG-систем, де безпека є емерджентною властивістю всієї системи, а не лише моделі.

Цей аналіз підводить до важливого висновку: сучасна безпека ШІ відходить від концепції «невразливої моделі» до концепції «стійкої до відмов, еластичної системи» (resilient AI system). Найефективнішою є стратегія захисту в глибину (defense-in-depth), де на кожному етапі обробки інформації – від отримання даних до генерації відповіді та взаємодії з інструментами – застосовуються незалежні рівні контролю. Модель є лише одним, хоч і центральним, компонентом у цьому складному архітектурному ланцюгу.

Висновки та перспективи подальших досліджень

Проведений аналіз сучасних методів захисту моделей штучного інтелекту від цілеспрямованих атак дозволяє зробити кілька ключових висновків.

По-перше, ландшафт загроз для систем ШІ зазнав значної еволюції. Якщо раніше основна увага була прикута до математичних вразливостей моделей, таких як чутливість до змагальних збурень, то з появою великомасштабних генеративних моделей, зокрема LLM, фокус змістився на атаки, що експлуатують логіку, контекст та взаємодію моделі з зовнішнім середовищем. Такі загрози, як ін'єкції запитів та маніпуляції в RAG-системах, розмивають традиційну межу між даними та інструкціями, перетворюючи будь-яке неперевірене джерело інформації на потенційний вектор атаки.

По-друге, найбільш критичними для сучасних систем є не ті атаки, що викликають миттєвий збій, а ті, що діють приховано, персистентно і підривають цілісність або конфіденційність на фундаментальному рівні. Атаки на ланцюг постачання, отруєння даних, викрадення моделей та витоки приватної інформації становлять найбільшу загрозу через складність їх виявлення, високу вартість усунення наслідків та потенціал для подальшої ескалації атак.

По-третє, у сфері захисних механізмів спостерігається чіткий розрив між формальними методами з доказовими гарантіями та практичними, архітектурними рішеннями. Хоча диференційна приватність та сертифікована стійкість пропонують надійний теоретичний захист, їхня висока обчислювальна вартість та негативний вплив на корисність моделі обмежують їх застосування для складних систем, як-от LLM. У результаті, на практиці перевага надається евристичним та інженерним підходам: посиленню запитів, впровадженню захисних бар'єрів (guardrails) та безпечному проєктуванню системних архітектур.

Насамкінець, найважливішим висновком є те, що ефективна безпека ШІ – це не властивість ізольованої моделі, а емерджентна характеристика всієї системи, в якій вона функціонує. Концепція «стійкої до відмов системи» (resilient system) витісняє ідею «невразливої моделі». Це вимагає застосування комплексного, багаторівневого підходу «захисту в глибину», де заходи безпеки впроваджуються на кожному етапі життєвого циклу – від верифікації даних до моніторингу взаємодії з користувачами та зовнішніми інструментами.

На основі виявлених проблем та невирішених питань можна окреслити кілька перспективних напрямів для майбутніх наукових досліджень:

1. Формальні гарантії для компонентів LLM-систем: замість спроб забезпечити сертифіковану стійкість для всієї монолітної LLM, що є практично неможливим, дослідження можуть зосередитися на розробці формальних методів для менших, але критично важливих компонентів системи. Наприклад, створення сертифіковано надійних моделей-класифікаторів для guardrails, які б гарантовано виявляли певні типи шкідливих запитів, або розробка верифікованих парсерів виводу, стійких до атак.

2. Стандартизація та бенчмарки для оцінювання безпеки LLM: наразі відсутні єдині, загальноприйняті стандарти та набори тестів (benchmarks) для кількісного оцінювання безпеки LLM-застосунків. Розробка таких стандартів, аналогічних існуючим у традиційній кібербезпеці, дозволить об'єктивно порівнювати стійкість різних моделей та захисних стратегій до таких атак, як ін'єкції запитів, витоки даних та отруєння.

3. Комбіновані та адаптивні захисти в RAG-ланцюгах: складні RAG-системи та ланцюги агентів (agent chains) потребують дослідження оптимальної композиції різних захисних механізмів. Майбутні роботи можуть вивчати синергетичний ефект від поєднання фільтрації даних на етапі пошуку, перевірки запитів перед подачею в модель та валідації виводу. Також перспективним є створення адаптивних систем захисту, які б динамічно змінювали рівень суворості контролю залежно від контексту діалогу та рівня довіри до джерел даних.

4. Методики системного моніторингу та виявлення аномалій: із зростанням складності та автономності ШІ-агентів виникає потреба в розробці масштабованих та ефективних рішень для моніторингу їхньої поведінки в реальному часі. Дослідження можуть бути спрямовані на створення методів, здатних виявляти складні, багатоетапні атаки, аналізуючи послідовності дій агента, його взаємодію з інструментами та інформаційні потоки, а не лише окремі запити та відповіді.

Ці напрями досліджень допоможуть подолати існуючий розрив між швидким розвитком можливостей ШІ та повільнішим прогресом у забезпеченні його безпеки, сприяючи створенню більш надійних та безпечних інтелектуальних систем майбутнього.

Література

[1]. Ilahi, I., Usama, M., Qamar, A. M., Al-Ghamdi, M., & Niyato, D. (2022). Challenges and Countermeasures for Adversarial Attacks on Deep Reinforcement Learning. *IEEE Transactions on Artificial Intelligence*, *3*(2), 90-109.

[2]. Liu, X., Xie, L., Wang, Y., Zou, J., Xiong, J., Ying, Z., & Vasilakos, A.V. (2021). Privacy and Security Issues in Deep Learning: A Survey. IEEE Access, 9, 4566-4593.

[3]. NIST. (2023, January 26). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. National Institute of Standards and Technology.

[4]. Cloudflare. (2023, October). *OWASP Top 10 risks for LLMs*. Cloudflare Learning Center.

[5]. OWASP. (2023). *OWASP Top 10 for Large Language Model Applications*.

[6]. Yuan, X., He, P., Zhu, Q., & Li, X. (2019). Adversarial Examples: Attacks and Defenses for Deep Learning. *IEEE Transactions on Neural Networks and Learning Systems*, *30*(9), 2805-2824.

[7]. Ponnuru, M. D. S., Amasala, L., Bhimavarapu, T. S., & Garikipati, G. C. (2023). A malware classification survey on adversarial attacks and defences. arXiv preprint arXiv:2312.09636.

[8]. Namatevs, I., Sudars, K., Nikulins, A., & Ozols, K. (2025). Privacy Auditing in Differential Private Machine Learning: The Current Trends. Applied Sciences, 15(2), 647.

[9]. Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S. Yu, and Xuyun Zhang. 2022. Membership Inference Attacks on Machine Learning: A Survey. ACM Comput. Surv. 54, 11s, Article 235.

[10]. Zhao, K., Li, L., Ding, K., Gong, N. Z., Zhao, Y., & Dong, Y. (2025). A Systematic Survey of Model Extraction Attacks and Defenses: State-of-the-Art and Perspectives. arXiv preprint arXiv:2508.15031.

[11]. Niu, J., Zhu, X., Zeng, M., Zhang, G., Zhao, Q., Huang, C., ... & Zhang, Y. (2025). Comparing Different Membership Inference Attacks with a Comprehensive Benchmark. IEEE Transactions on Information Forensics and Security.

[12]. Hu, H., & Pang, J. (2021). Stealing machine learning models: Attacks and countermeasures for generative adversarial networks. *Proceedings of the 37th Annual Computer Security Applications Conference*, 1-16.

[13]. Jagielski, M., Carlini, N., Berthelot, D., Kurakin, A., & Papernot, N. (2020). High accuracy and high fidelity extraction of neural networks. In 29th USENIX security symposium (USENIX Security 20) (pp. 1345-1362).

[14]. Zhou, Y., Ni, T., Lee, W. B., & Zhao, Q. (2025). A survey on backdoor threats in large language models (llms): Attacks, defenses, and evaluations. arXiv preprint arXiv:2502.05224.

[15]. Bagdasaryan, E., Poursaeed, O., & Shmatikov, V. (2019). Differential privacy has disparate impact on model accuracy. Advances in neural information processing systems, 32.

[16]. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 308–318.

[17]. Zhang, Qiuchen & Ma, Jing & Xiao, Yonghui & Lou, Jian & Xiong, Li. (2020). Broadening Differential Privacy for Deep Learning Against Model Inversion Attacks. 1061-1070. 10.1109/BigData50022.2020.9378274.

[18]. Li, L., Xie, T., & Li, B. (2023, May). Sok: Certified robustness for deep neural networks. In 2023 IEEE symposium on security and privacy (SP) (pp. 1289-1310). IEEE.

[19]. Wang, B., Jia, J., Cao, X., & Gong, N. Z. (2021, August). Certified robustness of graph neural networks against adversarial structural perturbation. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (pp. 1645-1653).

[20]. Kulynych, B., Hsu, H., Troncoso, C., & Calmon, F. P. (2023, June). Arbitrary decisions are a hidden cost of differentially private training. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (pp. 1609-1623).

[21]. Sharma, C. (2025). Retrieval-Augmented Generation: A Comprehensive Survey of Architectures, Enhancements, and Robustness Frontiers. arXiv preprint arXiv:2506.00054.

[22]. Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., & Ristenpart, T. (2016). Stealing machine learning models via prediction {APIs}. In *25th USENIX security symposium (USENIX Security 16)* (pp. 601-618).

[23]. Thudi, A., Jia, H., Meehan, C., Shumailov, I., & Papernot, N. (2024). Gradients look alike: Sensitivity is often

overestimated in {DP-SGD}. In 33rd USENIX Security Symposium (USENIX Security 24) (pp. 973-990).

[24]. Liu, Y., Wen, R., He, X., Salem, A., Zhang, Z., Backes, M., ... & Zhang, Y. (2022). {ML-Doctor}: Holistic risk assessment of inference attacks against machine learning models. In 31st USENIX Security Symposium (USENIX Security 22) (pp. 4525-4542)

[25]. Bulusu, S., Kailkhura, B., Li, B., Varshney, P. K., & Song, D. (2020). Anomalous example detection in deep learning: A survey. IEEE Access, 8, 132330-132347.

[26]. Denison, C., Ghazi, B., Kamath, P., Kumar, R., Manurangsi, P., Narra, K. G., ... & Zhang, C. (2022). Private ad modeling with DP-SGD. arXiv preprint arXiv:2211.11896.

[27]. Mao, Y., Balauca, S., & Vechev, M. (2024). Ctbench: A library and benchmark for certified training. arXiv preprint arXiv:2406.04848.

[28]. Liu, Y., Li, Z., Backes, M., Shen, Y., & Zhang, Y. (2023). Watermarking diffusion model. arXiv preprint arXiv:2305.12502

УДК 003.26:004.056.55:621.39(045)

Bondarovets S, Okhrimenko T. Modern artificial intelligence models protection methods from adversarial attacks

Abstract. *This article provides a comprehensive overview of the current state of artificial intelligence (AI) model security, systematizing vectors of targeted attacks and corresponding defense methods. The evolution of the threat landscape is analyzed, from classic attacks on machine learning (ML) models to specific vulnerabilities inherent in modern large language models (LLMs). The introduction outlines the problem's relevance in the context of AI's deep integration into critical infrastructure and business processes, emphasizing the shift from reactive vulnerability patching to proactive risk management, as reflected in industry standards like the NIST AI Risk Management Framework. The main body of the research begins with a detailed classification of modality-agnostic attacks, including adversarial examples, data poisoning, backdoors, model stealing, membership inference attacks, and model inversion. It then analyzes the most critical threats to modern systems based on criteria of prevalence, potential damage, and detection difficulty, focusing on supply chain attacks and data leakage. General-purpose defense methods are systematized according to the model lifecycle stages: at the data level (sanitization, differential privacy), during training (adversarial training, robust optimization), and at the inference stage (monitoring, security policies). A separate section is dedicated to the paradigm shift caused by LLMs. LLM-specific threats are examined in detail: prompt injection, including direct (jailbreaks) and indirect attacks in RAG systems, insecure output handling, and risks associated with fine-tuning and the use of external tools. Accordingly, multi-layered defense strategies for LLMs are analyzed, such as prompt hardening, the implementation of guardrails, red-teaming, and secure design of RAG systems and tools. An analytical synthesis summarizes the strengths and weaknesses of the reviewed approaches, assesses the risks of false positives and false negatives (FP/FN), and discusses their application conditions. The conclusion summarizes key findings and identifies promising directions for future research, including the development of formal security guarantees for LLMs, the standardization of benchmarks, and the advancement of comprehensive monitoring systems.*

Keywords*: artificial intelligence; cybersecurity; artificial intelligence protection; large language models; adversarial attacks.*

Бондаровець Сергій Сергійович, аспірант, Державний університет «Київський авіаційний інститут».
Bondarovets Serhii, PhD Student, State University «Kyiv Aviation Institute».

Охріменко Тетяна Олександрівна, кандидат технічних наук, старший дослідник, заступник декана з наукової роботи Факультету комп'ютерних наук та технологій, Державний університет «Київський авіаційний інститут».
Okhrimenko Tetiana, Ph.D., Senior Researcher, Deputy Dean for Scientific Work, Faculty of Computer Science and Technology, State University «Kyiv Aviation Institute».