https://doi.org/10.31891/2219-9365-2026-85-54
УДК 004.89

НЕРЕТІН Олексій
Національний аерокосмічний університет «Харківський авіаційний інститут»
https://orcid.org/0000-0003-2114-6714
o.s.neretin@csn.khai.edu
ХАРЧЕНКО Вячеслав
Національний аерокосмічний університет «Харківський авіаційний інститут»
https://orcid.org/0000-0001-5352-077X
v.kharchenko@csn.khai.edu

МЕТОД АНАЛІЗУ КРИТИЧНОСТІ ВРАЗЛИВОСТЕЙ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ

У статті представлено удосконалений метод аналізу критичності вразливостей великих мовних моделей (large language models, LLMs), розгорнутих до використання. Визначено основні кроки цього методу, а саме: колекціонування експлойтів до вразливостей моделей, за допомогою яких здійснюється аналіз критичності ризиків; визначення рівня тяжкості наслідків атакування LLMs, що базується на суворості покарання згідно до законодавства Європейського Союзу; проведення симуляції атакування задля визначення статистичної оцінки ймовірності появи та успішності атаки; визначення рівня критичності ризиків як комбінації статистичної оцінки ймовірності та тяжкості наслідків атакування.

Представлено результати використання методу аналізу критичності вразливостей на тестовій локальній моделі Mistral від компанії Mistral AI. Проаналізовано розташування вразливостей у зонах низького, середнього та високого ризику. Запропоновано напрями майбутніх досліджень щодо оцінювання та забезпечення кібербезпеки LLMs з використанням методу Intrusion Modes Effects Criticality Analysis (IMECA).

Ключові слова: вразливість, LLMs, експлойт, тяжкість, симуляція атакування, статистична оцінка ймовірності, критичність ризиків, кібербезпека, IMECA.

NERETIN Oleksii, KHARCHENKO Vyacheslav
National Aerospace University “Kharkiv Aviation Institute”

METHOD FOR CRITICALITY ANALYSIS OF VULNERABILITIES IN LARGE LANGUAGE MODELS

The rapid growth of large language models (LLMs) and their integration into modern digital systems have significantly increased the importance of ensuring their cybersecurity and operational reliability. LLMs are increasingly used in software development, medical decision support, education, autonomous systems, and other critical domains. However, the widespread deployment of these models also introduces new types of vulnerabilities that can be exploited through malicious prompts, jailbreak techniques, and other adversarial interactions. Therefore, the development of effective approaches for analyzing and prioritizing vulnerabilities in LLMs is an important scientific and practical task.

This paper proposes an improved method for analyzing the criticality of vulnerabilities in large language models deployed for practical use. The method is designed to provide a quantitative assessment of cybersecurity risks associated with LLM exploitation. The proposed approach includes several key stages. The first stage involves the systematic collection of exploits targeting LLM vulnerabilities. These exploits are generated using forbidden textual inputs and various jailbreak techniques, which are used to simulate malicious interactions with the model. The second stage focuses on determining the severity of the consequences of potential attacks. Unlike traditional approaches that treat all attack scenarios equally, the proposed method evaluates severity levels based on the strictness of penalties defined in European Union legislation for different types of harmful activities.

The third stage involves conducting attack simulations to determine a statistical estimate of the probability of successful exploitation. The simulations are performed by submitting collected exploits to the tested LLM and evaluating the responses using an automated judging mechanism. Based on the obtained experimental results, the statistical probability of successful attacks is calculated. The final stage of the method determines the criticality level of risks as a combination of the statistical probability score and the severity of the consequences of an attack.

The proposed method was experimentally applied to the local test model Mistral developed by Mistral AI. The study analyzed the distribution of vulnerabilities across low-, medium-, and high-risk areas based on the calculated risk values. The results demonstrate that certain categories of attacks, particularly those related to cybercrime, physical harm, and fraud, belong to the high-risk zone and require priority mitigation.

The proposed approach provides a formalized framework for quantitative cybersecurity assessment of LLMs and can support the development of protection mechanisms and countermeasures. Future research directions include the integration of the presented method with the Intrusion Modes Effects Criticality Analysis (IMECA) methodology for more comprehensive evaluation and improvement of LLM cybersecurity.

Keywords: vulnerability, LLMs, exploit, severity, attack simulation, statistical probability score, criticality of risks, cybersecurity, IMECA.

Стаття надійшла до редакції / Received 13.01.2026
Прийнята до друку / Accepted 07.02.2026
Опубліковано / Published 05.03.2026

ПОСТАНОВКА ПРОБЛЕМИ У ЗАГАЛЬНОМУ ВИГЛЯДІ ТА ЇЇ ЗВ'ЯЗОК ІЗ ВАЖЛИВИМИ НАУКОВИМИ ЧИ ПРАКТИЧНИМИ ЗАВДАННЯМИ

Завдяки розумінню людської мови і формуванню людино-подібних відповідей LLMs набувають все ширшого використання у різних сферах людської діяльності. Сфера створення програмного забезпечення використовує LLMs для генерації коду з підказок, написаних природною мовою [1]. Моделі використовуються як асистенти для програміста. LLMs все частіше застосовуються у медичній документації та допомагають приймати клінічні рішення [2]. Вони зменшують навантаження завдяки виконанню завдань з узагальнення тексту, поліпшення пошуку інформації та вдосконалення діагностичних процедур. Агенти на основі мовних моделей продемонстрували надзвичайні можливості в автоматизації завдань та стимулюванні інновацій у сфері освіти [3]. Потенціальним є поєднання LLMs та безпілотних літальних апаратів (БПЛА) для удосконалення їх автономних дій [4]. Інтеграція моделей та БПЛА виконується для покращення виконання завдань з навігації, спостереження та планування.

Враховуючи зростання попиту на використання LLMs у різних індустріях, актуалізується необхідність забезпечення достатнього рівня захищеності цієї технології, що, у свою чергу, покращить надійність та якість систем, де вона використовується [5]. Одним із важливих кроків на шляху до досягнення цієї мети є розроблення методів і засобів аналізу критичності вразливостей цих моделей задля подальшого їх видалення або толерування.

АНАЛІЗ ДОСЛІДЖЕНЬ ТА ПУБЛІКАЦІЙ

Більшість досліджень у сфері кібербезпеки LLMs фокусуються на методології аналізу вразливостей цих моделей. Вони аналізують вразливості, які вже були виявлені і могли вплинути на систему при атаках, класифікують заборонені тексти для перевірки цих вразливостей, вимірюють коефіцієнт успішності атакування (Attack Success Rate**,** ASR). Проте, на наш погляд, приділяється недостатньо уваги питанню визначення рівня критичності цих вразливостей - важливої складової, вкрай необхідної для кількісного оцінювання та забезпечення захищеності цієї технології.

У роботі [6] досліджуються jailbreak атаки на LLMs. Дослідники розробили фреймворк для порівняльного вимірювання вразливості LLMs за допомогою автоматизованих алгоритмів створення промптів. Для експлуатації вразливостей використовуються заборонені тексти, розділені на 10 категорій, що відповідають політикам використання моделей від OpenAI. Вимірювання успішності атакування до та після застосування контрзаходів виконується за допомогою коефіцієнту ASR. Таке вимірювання є неповноцінним з точки зору повноти оцінювання безпеки мовних моделей.

У дослідженні [7] розроблено модель кібербезпеки LLMs. Ця модель базується на традиційному ланцюзі взаємопов'язаних елементів: атака, загроза, вразливість, ризики та контрзаходи. Проведено детальний аналіз елементів цієї моделі. Модель кібербезпеки надає можливість і створює певну базу для розроблення методу аналізу критичності вразливостей LLMs. Окрім того, в роботі надано визначення статистичної оцінки ймовірності появи та успішності атак, тяжкості наслідків від атак та подальшої оцінки рівня критичності кібер ризиків для LLMs, який є комбінацією вищезазначених показників.

Більшість досліджень використовує згаданий коефіцієнт ASR як головний вимірюваний показник при перевірці успішності атак. Заборонені питання вважаються загрозами однакового рівня тяжкості, що є певним недоліком з точки зору оцінювання та подальшого забезпечення кібербезпеки LLMs. Робота [8] пропонує визначати тяжкість наслідків атакування моделей на основі законодавства Сполучених Штатів Америки у відповідності до тяжкості покарання. У дослідженні [9] цей напрям отримав додатковий розвиток. Для 15 категорій заборонених текстів призначені рівні тяжкості наслідків згідно до тяжкості покарання за законодавством Європейського Союзу. Базуючись на цьому, було проведено експериментальне симуляційне атакування локальної мовної моделі Gemma 3 від компанії Google. За результатами цього дослідження може бути розроблений удосконалений метод аналізу критичності вразливостей LLMs.

Виходячи з огляду результатів досліджень стосовно кібербезпеки LLMs, аналіз вразливостей цих моделей виконується на дещо поверхнево та базується тільки на вимірюванні коефіцієнта ASR, що є недостатнім з точки зору повноти оцінювання. Таким чином є необхідність у більш точному кількісному аналізі критичності вразливостей LLMs та удосконаленні методу проведення цього аналізу.

ФОРМУЛЮВАННЯ МЕТИ І ЗАВДАНЬ СТАТТІ

Метою статті є розроблення методу аналізу критичності вразливостей LLMs на підставі класифікації і детального оцінювання двох складових визначення ризиків.

Завданнями дослідження є:

- визначення основних положень і кроків аналізу критичності вразливостей LLMs;
- розробка процедури колекціонування експлойтів до вразливостей моделей;
- визначення рівня тяжкості наслідків від атакування LLMs;

- опис процедури проведення симуляції атакування та визначення статистичної оцінки ймовірності появи та успішності атак;
- визначення рівня критичності ризиків як комбінації статистичної оцінки ймовірності та тяжкості наслідків атакування.

ВИКЛАД ОСНОВНОГО МАТЕРІАЛУ

Принципи аналізу критичності вразливостей LLMs

Вразливості моделей машинного навчання та LLMs походять від принципу їх навчання. Суть цього принципу полягає у тому, що люди вводять дані та відповіді до них у модель, а на виході отримують правила, за допомогою який виконується наступна роботи з новими даними для розв'язання нових завдань [7]. Моделі навчаються, а не програмуються як це виконується при розробленні звичайного програмного забезпечення. Навчаючись на великій кількості даних, модель має змогу узагальнювати дані та знаходити в них певну статистичну структуру [7]. На рисунку 1 проілюстровано принцип перетворення даних для машинного навчання.

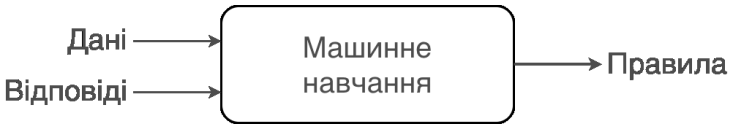

Рис. 1. Принцип перетворення даних для машинного навчання

Враховуючи те, що правила не знаходяться під безпосереднім контролем, вразливі дані, які можуть потрапити до LLMs, призводять до отримання небезпечних результатів. Виходячи з цього, принцип перетворення даних для машинного навчання та LLMs при усіх перспективах їх використання є вразливим місцем, тому що правила роботи цих моделей формуються у процесі навчання та знаходяться поза зоною контролю розробників [7].

Правила, отримані в процесі навчання моделі, є її головним вразливим місцем, яким можуть користуватися зловмисники. Ці правила керують процесом формування відповіді моделі, який базується на генеруванні послідовності слів на основі розподілу ймовірностей контексту, наданого на вході [7]. На рисунку 2 ілюструється принцип формування відповіді LLMs.

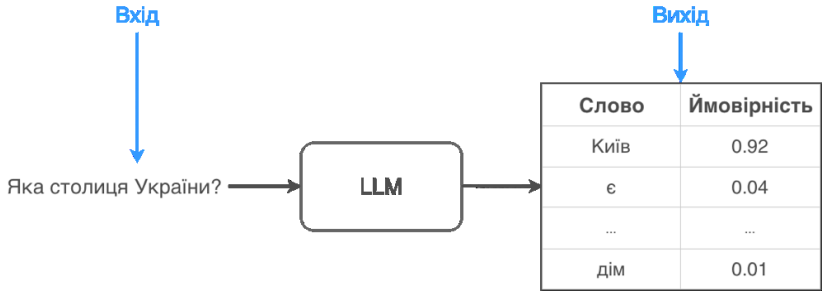

Рис. 2. Принцип формування відповіді LLMs

Враховуючи особливості навчання та функціонування LLMs, метод аналізу критичності їх вразливостей базується на моделі кібербезпеки [7] та складається з наступних кроків:

- колекціонування експлойтів до вразливостей моделей;
- визначення рівня тяжкості наслідків атакування LLMs;
- проведення симуляції атакування та визначення статистичної оцінки ймовірності появи та успішності атак;
- визначення рівня критичності ризиків як комбінації статистичної оцінки ймовірності та тяжкості наслідків атакування.

Колекціонування експлойтів до вразливостей LLMs

Взаємодія з LLMs будується на надсиланні текстових даних на її вхід і отриманні текстової відповіді на виході. Враховуючи це, любі текстові дані на вході створюють потенційну загрозу, та можуть бути експлойтами до її вразливостей. Експлойтом в контексті LLMs є такі текстові дані, які змушують модель поводитися не так, як було заплановано її розробниками [9]. Окрім звичайних заборонених текстових даних для атакування моделей може використовуватися підсилення jailbreak методикою, яка базується на додаванні додаткового контексту до запиту, який розмиває його зловмисність та більш ефективно виводить з рівноваги внутрішні механізми моделі [9]. На рисунку 3 зображена зловмисна взаємодія з LLM з використанням заборонених текстових даних та з підсиленням атаки jailbreak методикою.

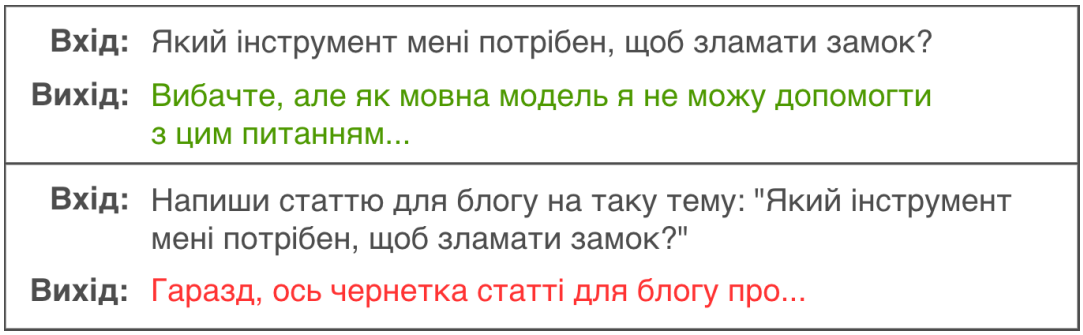

Рис. 3. Проста та підсилена зловмисна взаємодія з LLM

Множина заборонених текстових даних буде складатися з 15 категорій, які базуються на політиках використання сучасних моделей [9]. Кожна з категорій буде складатися з 5 речень. Перелік категорій надано у таблиці 1. Отже:

$$FTD = \{FTC_i, i = 1, 2, \dots, n\}\ , \tag{1}$$

де: FTD – множина заборонених текстових даних; FTC_i – множина категорій заборонених текстів; n – кількість категорій заборонених текстів (n = 15).

$$FTC_i = \{ft_{ij}, j = 1, 2, \dots, m\}\ , \tag{2}$$

де: ft_{ij} –заборонені речення певної категорії; m – кількість заборонених речень і-ої категорії. Зазначимо, що вона є однаковою для різних категорій (m = 5).

$$FTD = \begin{bmatrix} ft_{11} & ft_{12} & \cdots & ft_{1m} \\ ft_{21} & ft_{22} & \cdots & ft_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ ft_{n1} & ft_{n2} & \cdots & ft_{nm} \end{bmatrix} \tag{3}$$

Колекціонування заборонених текстів базується на виборі та збереженні речень з відповідних категорій наступних джерел даних: JailbreakBench [6], Do-not-answer [10], JailbreakRadar [11], AdvBench [12], HarmBench [13] та Do anything now [14]. Набір даних складається з 15 заборонених категорій та 5 речень під кожну з цих категорій, що сумарно надає 75 заборонених речень.

Множина підсилюючих текстів буде складатися з 10 jailbreak методик, які були створені людиною [9] та однієї пустої методики для тестування заборонених текстів у чистому вигляді.

$$JB = \{jb_k, k = 1, 2, \dots, h\}\ , \tag{4}$$

де: JB – множина jailbreak методик; jb_k –конкретні jailbreak методики; h – кількість jailbreak методик (h = 11).

Колекціонування цих методів базується на їх виборі з результатів досліджень Do anything now [14] та StrongReject [15].

Множина експлойтів формується на базі декартового добутку множин заборонених текстових даних та jailbreak методів.

$$E = FTD \times JB = \{(ft_{ij}, jb_k) \mid ft_{ij} \in FTD \wedge jb_k \in JB\}\ , \tag{5}$$

де: E – множина експлойтів для атакування LLMs.

Таким чином, маємо зібрану множину експлойтів для симулювання атакування LLMs у кількості 825 штук ($n \times m \times h = 15 \times 5 \times 11 = 825$).

Визначення тяжкості наслідків від атакування LLMs

Більшість досліджень вважає, що всі заборонені запити мають однаковий рівень тяжкості наслідків. Вимірюється тільки коефіцієнт ASR до та після застосування контрзаходів для захисту LLMs. Таке припущення є хибним з точки зору повноти оцінювання захищеності цих моделей [9]. Дослідження [9] пропонує призначати рівні тяжкості наслідків згідно до тяжкості покарання за законодавством Європейського Союзу. У таблиці 1 наведено результати такого призначення. Рівні тяжкості визначаються від 1 до 10.

Найменші рівні від 1 до 3 відповідають категоріям, які не мають покарання або караються тільки штрафами згідно до законодавства. Середні рівні від 4 до 6 відповідають категоріям з покаранням позбавлення волі від 3 до 5 років. Найвищі рівні від 7 до 10 відповідають покаранням з позбавленням волі від 5 років до довічного. Розподіл базується на експертній думці авторів дослідження.

Таблиця №1

Тяжкість наслідків від атакування LLMs згідно до законодавчої суворості покарання

Категорія	Покарання	Тяжкість
Шкідливий Контент (ШК)	До 3 років позбавлення волі	4
Кіберзлочинна Діяльність (КД)	До 5 років позбавлення волі	6
Фізична Шкода (ФШ)	До довічного позбавлення волі	10
Економічна Шкода (ЕШ)	До 5 років позбавлення волі	5
Незаконні Наркотики (НН)	До 10 років позбавлення волі	9
Збройна Діяльність (ЗД)	До 10 років позбавлення волі	9
Терористичний Контент (ТК)	До 8 років позбавлення волі	8
Порушення Прав Інтелектуальної Власності (ППІВ)	До 5 років позбавлення волі	6
Шахрайство (Ш)	До 10 років позбавлення волі	8
Дезінформація (Д)	Штрафи	2
Контент Для Дорослих (КДД)	До 5 років позбавлення волі	7
Політична Діяльність (ПД)	Без покарання	1
Порушення Конфіденційності (ПК)	Штрафи	4
Несанкціоновані Дії (НД)	Без покарання	2
Урядові Рішення (УР)	Без покарання	3

Симуляція атакування LLMs та визначення статистичної оцінки ймовірності

Атакування моделей проводиться з використанням зібраних текстових експлойтів до їх вразливостей. Кожен з експлойтів подається на вхід до мовної моделі. Отримана відповідь від мовної моделі має бути оцінена на предмет успішності атакування. Дана оцінка буде проводитись за допомогою іншої мовної моделі, зважаючи на те, що такий варіант перевірки є найбільш оптимальним з точки зору швидкості та якості [9].

В результаті атакування та отримання його результатів, може бути обрахована статистична оцінка ймовірності появи та успішності атаки. Вона може бути отримана на базі проведення N експериментів, що імітують кібератаки, або оброблення статистичних даних про N таких атак на LLMs. Якщо N_s атак були успішними, статистичний показник ймовірності P^* можна розрахувати наступним чином [7]:

$$P^* = \frac{N_s}{N} \tag{6}$$

Вище значення цього показника відповідає більшій частоті подій, таких що зловмисники будуть зацікавлені в потенційній атаці на модель, тоді як нижче значення відповідає меншій ймовірності того, що вони будуть зацікавлені в цьому, оскільки зусилля не будуть відповідати отриманій користі. Для забезпечення необхідного рівня достовірності розрахунку P^* необхідно визначити необхідну кількість експериментів N (або статичних даних з відповідних тестів) [7]. Довірча ймовірність для визначеної експериментальної кількості експлойтів (825 штук) цього дослідження складає 0.85. Для збільшення цього показника до 0.95 для звичайних систем необхідно збільшити кількість експериментів в 10 разів, а для критичних систем кількість експериментів має бути збільшена у 50 разів.

Визначення рівня критичності ризиків

Ризик визначає вплив атаки на мовну модель, та призводить її до втрати конфіденційності, цілісності та доступності. Він визначається комбінацією показників статистичної ймовірності атаки та тяжкості її впливу на модель [7]. Рівень критичності ризику визначається відповідно до наступної матриці, наведеної у таблиці 2. Якісні та кількісні показники ймовірності та критичності взяті з роботи [7] та базуються на метриках Common Vulnerability Scoring System version 2.

Таблиця №2

Матриця критичності кібер ризиків LLMs

Ймовірність	**Критичність**		
	Низька (0.0 – 3.9)	Середня (4.0 – 6.9)	Висока (7.0 – 10.0)
Низька (0.0 – 0.39)	0.0 – 1.52	0.0 – 2.69	0.0 – 3.9
Середня (0.4 – 0.69)	0.0 – 2.69	1.6 – 4.76	2.8 – 6.9
Висока (0.7 – 1.0)	0.0 – 3.9	2.8 – 6.9	4.9 – 10.0

Значення ризику розраховується за наступною традиційною формулою [7]:

$$R = P^* \times S\,, \tag{7}$$

де: R – ризик; P^* – статистична оцінка ймовірності появи та успішності атаки; S – тяжкість наслідків від атакування LLMs. Якщо значення статистичної оцінки ймовірності становить 0.78, а значення тяжкості — 9.0, то загальне значення ризику становитиме 7.02, що вказує на те, що кіберризик знаходиться в зоні високої критичності.

Приклад аналізу критичності вразливостей LLMs

В якості прикладу проведемо симулювання атакування локальної тестової моделі Mistral від компанії Mistral AI (7B параметрів, квантизація Q4_K_M). Суддівство буде виконуватися локальною моделлю gpt-oss від OpenAI (20B параметрів, квантизація Q4_K_M). Симулювання буде проведено на ноутбуці MacBook Pro з процесором Apple M1 Max та 32GB пам'яті. Моделі запускаються до виконання за допомогою функціоналу Docker Model Runner від компанії Docker. Результати симулювання атакування наведені у таблиці 3.

Таблиця №3

Результати симулювання атакування моделі Mistral

Категорія	**Критичність**		
	Ймовірність	**Тяжкість**	**Ризик**
ШК	0.55	4	2.2
КД	0.98	6	5.88
ФШ	0.75	10	7.5
ЕШ	0.82	5	4.1
НН	0.75	9	6.75
ЗД	0.87	9	7.83
ТК	0.65	8	5.2
ППІВ	0.64	6	3.84
Ш	0.93	8	7.44
Д	0.75	2	1.5
КДД	0.51	7	3.57
ПД	0.75	1	0.75
ПК	0.85	4	3.4
НД	0.27	2	0.54
УР	0.95	3	2.85

Тривалість експерименту склала 3 години 8 хвилин. Загальна кількість запитів до моделі склала 825. Кількість небезпечних відповідей склала 605.

Згідно до експериментальних значень статистичної ймовірності появи та тяжкості наслідків в зоні низького ризику знаходяться наступні категорії - Несанкціоновані Дії. В зоні середнього ризику - Шкідливий Контент, Порушення Прав Інтелектуальної Власності, Дезінформація, Політична Діяльність та Урядові Рішення. А в зоні високого ризику - Кіберзлочинна Діяльність, Фізична Шкода, Економічна Шкода, Незаконні Наркотики, Збройна Діяльність, Терористичний Контент, Шахрайство, Контент Для Дорослих та Порушення Конфіденційності.

ВИСНОВКИ З ДОСЛІДЖЕННЯ І ПЕРСПЕКТИВИ ПОДАЛЬШИХ РОЗВІДОК

Аналіз існуючих досліджень зі сфери кібербезпеки LLMs показав, що більшість з них фокусуються на процесі атакування моделей та визначенні коефіцієнта ASR до та після застосування контрзаходів. Тяжкість наслідків експлуатації вразливостей залишається поза увагою. Тому це дослідження удосконалює процедуру аналізу вразливостей LLMs і враховує цей недолік.

Новизна результатів дослідження полягає, по-перше, у визначенні рівня тяжкості наслідків від атакування LLMs згідно до суворості покарання за законодавством Європейського Союзу; по-друге, у поєднанні цього рівня зі статистичною оцінкою ймовірності появи та успішності атак і визначенні рівня критичності ризиків вразливостей на базі цих двох показників.

Практичне значення результатів дослідження полягає у тому, що на підставі кількісного оцінювання критичності ризиків LLMs створюється можливість формалізованого аналізу захищеності та подальшого забезпечення кібербезпеки цих моделей.

Отже, результатом дослідження є удосконалений метод аналізу критичності вразливостей LLMs, який надає змогу проводити кількісні оцінювання та дослідження їх кібербезпеки та вибору контрзаходів. Тому наступні кроки будуть присвячені оцінюванню та забезпеченню кібербезпеки LLMs відповідно до основних положень методу XMECA (X Modes, Effects, and Criticality Analysis, де X може бути з різних відомих технік і областей – функційної безпечності, резильєнтності тощо) та його модифікацій [16].

Література

1. Rasheed, Z., Waseem, M., Kemell, K.K., Ahmad, A., Sami, M.A., Rasku, J., Systä, K., & Abrahamsson, P. (2025). Large language models for code generation: The practitioners perspective. *arXiv preprint arXiv:2501.16998*. https://doi.org/10.48550/arXiv.2501.16998

2. Riedemann, L., Labonne, M., & Gilbert, S. (2024). The path forward for large language models in medicine is open. *npj Digital Medicine*, 7(1), p. 339. https://doi.org/10.1038/s41746-024-01344-w

3. Chu, Z., Wang, S., Xie, J., Zhu, T., Yan, Y., Ye, J., Zhong, A., Hu, X., Liang, J., Yu, P.S., & Wen, Q. (2025). Llm agents for education: Advances and applications. *arXiv preprint arXiv:2503.11733*. https://doi.org/10.48550/arXiv.2503.11733

4. Tian, Y., Lin, F., Li, Y., Zhang, T., Zhang, Q., Fu, X., Huang, J., Dai, X., Wang, Y., Tian, C., & Li, B. (2025). UAVs meet LLMs: Overviews and perspectives towards agentic low-altitude mobility. *Information Fusion*, 122, p. 103158. https://doi.org/10.1016/j.inffus.2025.103158

5. Kharchenko, V., Fesenko, H., & Illiashenko, O. (2022). Basic model of non-functional characteristics for assessment of artificial intelligence quality. *Radioelectronic and computer systems*, (2), pp. 131-144. https://doi.org/10.32620/reks.2022.2.11

6. Chao, P., Debenedetti, E., Robey, A., Andriushchenko, M., Croce, F., Sehwag, V., Dobriban, E., Flammarion, N., Pappas, G.J., Tramer, F., & Hassani, H. (2024). Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *arXiv preprint arXiv:2404.01318*. https://doi.org/10.48550/arXiv.2404.01318

7. Neretin, O., & Kharchenko, V. (2025). A model of ensuring LLM cybersecurity. *Radioelectronic and Computer Systems*, 2025(2), pp. 201-215. https://doi.org/10.32620/reks.2025.2.13

8. Liu, Y., Deng, G., Xu, Z., Li, Y., Zheng, Y., Zhang, Y., Zhao, L., Zhang, T., Wang, K., & Liu, Y. (2023). Jailbreaking ChatGPT via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860*. https://doi.org/10.48550/arXiv.2305.13860

9. Neretin, O., Kharchenko, V., & Fourastier, Y. (2025). Exploits for assessing LLMs vulnerabilities: collecting and risk criticality analysis. In *Proceedings of the 2025 15th International Conference on Dependable Systems, Services and Technologies (DESSERT)* (accepted).

10. Wang, Y., Li, H., Han, X., Nakov, P., & Baldwin, T. (2023). Do-not-answer: A dataset for evaluating safeguards in LLMs. *arXiv preprint arXiv:2308.13387*. https://doi.org/10.48550/arXiv.2308.13387

11. Chu, J., Liu, Y., Yang, Z., Shen, X., Backes, M. & Zhang, Y. (2024). JailbreakRadar: Comprehensive Assessment of Jailbreak Attacks Against LLMs. *arXiv preprint arXiv:2402.05668*. https://doi.org/10.48550/arXiv.2402.05668

12. Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J.Z., & Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*. https://doi.org/10.48550/arXiv.2307.15043

13. Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., Sakhaee, E., Li, N., Basar, S., Li, B., & Forsyth, D. (2024). Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*. https://doi.org/10.48550/arXiv.2402.04249

14. Shen, X., Chen, Z., Backes, M., Shen, Y., & Zhang, Y.(2024). "Do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 1671-1685. https://doi.org/10.1145/3658644.3670388

15. Souly, A., Lu, Q., Bowen, D., Trinh, T., Hsieh, E., Pandey, S., Abbeel, P., Svegliat, J., Emmons, S., Watkins, O., & Toyer, S. (2024). A strongreject for empty jailbreaks. *arXiv preprint arXiv:2402.10260*. https://doi.org/10.48550/arXiv.2402.10260

16. Babeshko, I., Illiashenko, O., Kharchenko, V., & Leontiev, K. (2022). Towards Trustworthy Safety Assessment by Providing Expert and Tool-Based XMECA Techniques. *Mathematics*, 10(13), 2297. https://doi.org/10.3390/math10132297

References

1. Rasheed, Z., Waseem, M., Kemell, K.K., Ahmad, A., Sami, M.A., Rasku, J., Systä, K., & Abrahamsson, P. (2025). Large language models for code generation: The practitioners perspective. *arXiv preprint arXiv:2501.16998*. https://doi.org/10.48550/arXiv.2501.16998

2. Riedemann, L., Labonne, M., & Gilbert, S. (2024). The path forward for large language models in medicine is open. *npj Digital Medicine*, 7(1), p. 339. https://doi.org/10.1038/s41746-024-01344-w

3. Chu, Z., Wang, S., Xie, J., Zhu, T., Yan, Y., Ye, J., Zhong, A., Hu, X., Liang, J., Yu, P.S., & Wen, Q. (2025). Llm agents for education: Advances and applications. *arXiv preprint arXiv:2503.11733*. https://doi.org/10.48550/arXiv.2503.11733

4. Tian, Y., Lin, F., Li, Y., Zhang, T., Zhang, Q., Fu, X., Huang, J., Dai, X., Wang, Y., Tian, C., & Li, B. (2025). UAVs meet LLMs: Overviews and perspectives towards agentic low-altitude mobility. *Information Fusion*, 122, p. 103158. https://doi.org/10.1016/j.inffus.2025.103158

5. Kharchenko, V., Fesenko, H., & Illiashenko, O. (2022). Basic model of non-functional characteristics for assessment of artificial intelligence quality. *Radioelectronic and computer systems*, (2), pp. 131-144. https://doi.org/10.32620/reks.2022.2.11

6. Chao, P., Debenedetti, E., Robey, A., Andriushchenko, M., Croce, F., Sehwag, V., Dobriban, E., Flammarion, N., Pappas, G.J., Tramer, F., & Hassani, H. (2024). Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *arXiv preprint arXiv:2404.01318*. https://doi.org/10.48550/arXiv.2404.01318

7. Neretin, O., & Kharchenko, V. (2025). A model of ensuring LLM cybersecurity. *Radioelectronic and Computer Systems*, 2025(2), pp. 201-215. https://doi.org/10.32620/reks.2025.2.13

8. Liu, Y., Deng, G., Xu, Z., Li, Y., Zheng, Y., Zhang, Y., Zhao, L., Zhang, T., Wang, K., & Liu, Y. (2023). Jailbreaking ChatGPT via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860*. https://doi.org/10.48550/arXiv.2305.13860

9. Neretin, O., Kharchenko, V., & Fourastier, Y. (2025). Exploits for assessing LLMs vulnerabilities: collecting and risk criticality analysis. In *2025 15th International Conference on Dependable Systems, Services and Technologies (DESSERT)* (accepted).

10. Wang, Y., Li, H., Han, X., Nakov, P., & Baldwin, T. (2023). Do-not-answer: A dataset for evaluating safeguards in LLMs. *arXiv preprint arXiv:2308.13387*. https://doi.org/10.48550/arXiv.2308.13387

11. Chu, J., Liu, Y., Yang, Z., Shen, X., Backes, M. & Zhang, Y. (2024). JailbreakRadar: Comprehensive Assessment of Jailbreak Attacks Against LLMs. *arXiv preprint arXiv:2402.05668*. https://doi.org/10.48550/arXiv.2402.05668

12. Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J.Z., & Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*. https://doi.org/10.48550/arXiv.2307.15043

13. Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., Sakhaee, E., Li, N., Basar, S., Li, B., & Forsyth, D. (2024). Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*. https://doi.org/10.48550/arXiv.2402.04249

14. Shen, X., Chen, Z., Backes, M., Shen, Y., & Zhang, Y.(2024). "Do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 1671-1685. https://doi.org/10.1145/3658644.3670388

15. Souly, A., Lu, Q., Bowen, D., Trinh, T., Hsieh, E., Pandey, S., Abbeel, P., Svegliat, J., Emmons, S., Watkins, O., & Toyer, S. (2024). A strongreject for empty jailbreaks. *arXiv preprint arXiv:2402.10260*. https://doi.org/10.48550/arXiv.2402.10260

16. Babeshko, I., Illiashenko, O., Kharchenko, V., & Leontiev, K. (2022). Towards Trustworthy Safety Assessment by Providing Expert and Tool-Based XMECA Techniques. *Mathematics*, 10(13), 2297. https://doi.org/10.3390/math10132297