КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 1 (29), 2025
ISSN 2663 - 4023
DOI 10.28925/2663-4023.2025.29.918
УДК 004.89

Гайдур Галина Іванівна
доктор технічних наук, професор, завідувач кафедри систем та технологій кібербезпеки
Державний університет інформаційно-телекомунікаційних технологій, Київ, Україна
ORCID ID: 0000-0003-0591-3290
g.haidur@duikt.edu.ua

Власенко Вадим Олександрович
кандидат технічних наук, доцент, доцент кафедри систем та технологій кібербезпеки
Державний університет інформаційно-телекомунікаційних технологій, Київ, Україна
ORCID ID: 0000-0002-9329-5914
v.vlasenko@duikt.edu.ua

Петрова Олександра Сергіївна
студентка кафедри систем та технологій кібербезпеки,
Державний університет інформаційно-телекомунікаційних технологій, Київ, Україна
ORCID ID: 0009-0008-1535-9215
pet.85saf@gmail.com

ЗАХИСТ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ: РИЗИКИ, ЗАГРОЗИ ТА ПІДХОДИ ДО БЕЗПЕКИ

Анотація. У статті здійснено комплексний аналіз сучасних викликів у сфері безпеки великих мовних моделей (Large Language Models, LLM), які стали ключовим елементом цифрової трансформації у багатьох галузях. Розглянуто характерні загрози, що виникають як унаслідок цілеспрямованих атак на моделі, так і через їхнє зловмисне використання в кіберзлочинності. Визначено основні вектори ризику, серед яких найбільш небезпечними є prompt injection — впровадження у запит прихованих інструкцій для зміни логіки роботи моделі, а також джейлбрейкінг — формування запитів, що дозволяють обійти вбудовані обмеження та ініціювати небажану поведінку. Окремо підкреслено ризики витоку конфіденційних даних з навчальних наборів, генерації шкідливого або вразливого коду, здатного потрапити у виробничі середовища, а також поширення дезінформаційного контенту, включно з мультимедійними матеріалами типу deepfake. На основі проведеного аналізу запропоновано концептуальну модель безпеки LLM, яка передбачає поєднання технічних, архітектурних і нормативно-правових елементів захисту. Значна увага приділена оцінці та практичному застосуванню таких механізмів, як AI-фаєрволи — проміжні системи, що перевіряють запити й відповіді моделі; вбудовані захисні модулі, інтегровані в архітектуру LLM; а також guardrails — обмеження вихідних даних без втручання в параметри моделі. Додатково розглянуто методи watermarking для ідентифікації синтетичного контенту та інструменти виявлення згенерованої інформації. Важливою складовою є нормативне регулювання, яке встановлює рамки використання потужних моделей і створює основу для зниження ризиків зловживання. Зроблено висновок, що традиційні засоби кіберзахисту, орієнтовані на статичні або сигнатурні методи виявлення, є недостатніми для генеративних систем, які функціонують у динамічному мовному середовищі. Для підвищення рівня безпеки необхідно впроваджувати багаторівневі стратегії, що охоплюють усі етапи життєвого циклу LLM — від розробки й навчання до практичного застосування та регуляторного контролю.

Ключові слова: мовні моделі; Generative AI; кібербезпека; AI-фаєрвол; prompt injection; guardrails; watermarking; вразливість LLM.

ВСТУП

Розвиток генеративних технологій штучного інтелекту, зокрема великих мовних моделей (Large Language Models, LLM), таких як GPT, Claude та PaLM, відкриває значні

можливості для автоматизації обробки природної мови, створення креативного контенту, аналізу даних та підтримки прийняття рішень. Однак, одночасно з інноваційним потенціалом, LLM формують новий спектр безпекових викликів, що суттєво відрізняються від традиційних кіберзагроз.

Постановка проблеми. На відміну від класичних програмних систем, LLM є висококомплексними стохастичними моделями, схильними до непередбачуваної поведінки та здатними генерувати контент, який може бути шкідливим, токсичним або містити конфіденційну інформацію. Серед найбільш небезпечних векторів атак виділяються prompt injection, джейлбрейкінг, генерація шкідливого коду та зловживання інтеграціями з API. Ці ризики посилюються відсутністю універсальних стандартів безпеки та високим рівнем інтеграції LLM у критичні бізнес-процеси.

Аналіз останніх досліджень та публікацій. Аналіз сучасних наукових досліджень та галузевих звітів [1]–[8] показує, що питання безпеки великих мовних моделей (LLM) стає одним із ключових напрямів досліджень у сфері кібербезпеки та штучного інтелекту. В останні роки різко зросла кількість публікацій, присвячених як технічним, так і етичним аспектам захисту генеративних моделей.

OpenAI у своїй системній карті GPT-4 [1] наводить перелік ключових ризиків, серед яких — prompt injection, джейлбрейкінг та витоки конфіденційних даних із навчальних наборів. Вони підкреслюють, що через складність моделі навіть добре налаштовані фільтри можуть бути обійдені за допомогою обфускації інструкцій або контекстної інженерії запитів. Наводяться приклади атак, де зловмисники, модифікуючи вхідний prompt, отримували доступ до функцій, які повинні бути заблокованими.

Carlini та ін. [2] у роботі *Extracting Training Data from Large Language Models* продемонстрували експериментально, що LLM здатні відтворювати частини своїх навчальних даних, навіть якщо ці дані були унікальними та зустрічались у навчальному корпусі лише один раз. Це створює критичний ризик для захисту персональних даних, особливо у випадках, коли моделі навчались на великих обсягах неанонімізованої інформації.

Weidinger та ін. [3] акцентують на соціальних і етичних наслідках використання LLM, таких як поширення упередженої, дискримінаційної або токсичної інформації. У статті розглядаються приклади, коли моделі відтворювали расистські або сексистські висловлювання, а також описується вплив цих проблем на користувачів і суспільство в цілому.

Bender та ін**.** [4] у роботі *On the Dangers of Stochastic Parrots* ввели термін «стохастичні папуги» для опису LLM, які без розуміння відтворюють патерни зі своїх навчальних даних. Автори вказують на ризик, що такі моделі можуть масово поширювати дезінформацію або повторювати помилкові відомості, якщо вони закладені у навчальних корпусах.

Microsoft [6] у документі *Responsible AI Standard v2* пропонує багаторівневий підхід до безпеки LLM, включаючи механізми керування ризиками, моніторинг використання та обмеження доступу до певних функцій. Водночас вони наголошують, що навіть при наявності таких механізмів LLM залишаються вразливими до атак через інтеграцію з API та сторонніми сервісами.

NIST у *AI Risk Management Framework (AI RMF 1.0)* [8] формулює підхід до управління ризиками у використанні LLM, підкреслюючи необхідність розгляду безпеки на всіх етапах життєвого циклу — від збирання даних до розгортання та експлуатації моделі.

Важливою проблемою, що простежується у більшості робіт, є обмеженість класичних методів кіберзахисту (контроль доступу, сигнатурна фільтрація, sandboxing) у застосуванні до генеративних моделей.

Мета статті. Мета полягає у створенні системного аналізу актуальних безпекових викликів, пов'язаних із використанням великих мовних моделей (LLM), та розробка концептуальної моделі їх захисту, що поєднує технічні, архітектурні та нормативно-правові рішення.

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Унікальні загрози безпеці Generative AI

Попри унікальні можливості, моделі Generative AI залишаються вразливими до навмисних атак і маніпуляцій з боку зловмисників. Серед найбільш небезпечних загроз — джейлбрейкінг та prompt injection, які становлять серйозний ризик як для самих моделей, так і для застосунків, що їх використовують.

Джейлбрейкінг — це відносно нова техніка, коли спеціально сформовані запити змушують модель штучного інтелекту генерувати небезпечний, некоректний або оманливий контент. Унаслідок цього модель може ігнорувати власні протоколи безпеки або етичні обмеження. За аналогією зі смартфонами, де джейлбрейкінг відкриває прихований доступ до системи, в AI-середовищі це означає обхід внутрішніх механізмів безпеки з метою створення забороненого чи небажаного контенту.

Атаки типу prompt injection полягають у впровадженні в потік вхідних даних шкідливих інструкцій або даних, що змушують модель виконувати команди зловмисника замість завдань, визначених розробником. Механізм подібний до SQL-ін'єкцій у базах даних, коли шкідливий код у запиті інтерпретується системою як нова команда. У контексті Generative AI такі ін'єкції можуть змінювати роботу моделі, створюючи вивід, який значно відрізняється від початкового призначення програми. Особливо небезпечно це у випадках, коли модель має доступ до зовнішніх інструментів, плагінів або API, що значно розширює поверхню атаки.

На рис. 1 наведено порівняння двох поширених типів атак на великі мовні моделі. Prompt injection передбачає додавання до вхідних даних прихованих або навмисно сформованих інструкцій, що змінюють системні правила роботи моделі та призводять до небажаної поведінки (наприклад, розкриття внутрішніх інструкцій чи виконання забороненого завдання). Jailbreak полягає у створенні спеціальних запитів, метою яких є обхід вбудованих механізмів безпеки або етичних обмежень, унаслідок чого модель може згенерувати шкідливу, заборонену чи неетичну відповідь.

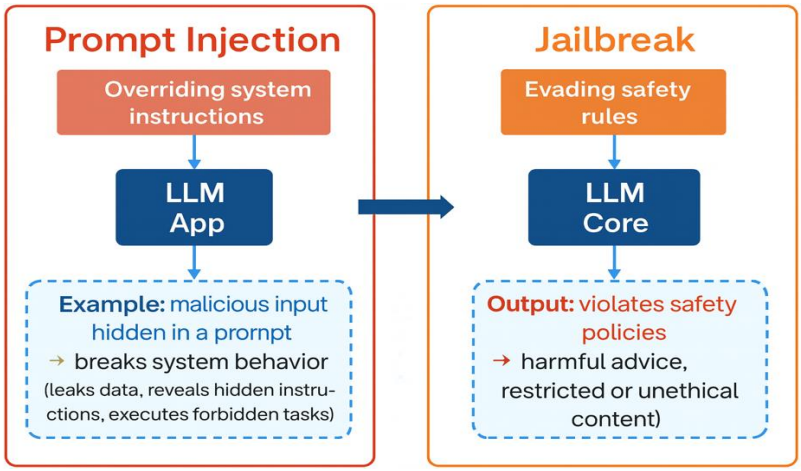

Рис. 1. Схематичне порівняння атак Prompt Injection та Jailbreak у LLM-системах

ISSN 2663 - 4023

Наслідки обох атак можуть бути критичними: від репутаційних збитків, якщо модель згенерує образливий чи абсурдний контент, до реальних інцидентів кібербезпеки, коли LLM використовується у пошуку, роботизованих системах або програмах з можливістю виконувати дії у реальному світі.

На рис. 2 наведено приклад prompt injection, що ілюструє потенційні ризики інтеграції LLM у комерційні сервіси без належних механізмів валідації вводу. Зловмисник (або жартівник) додав у діалог інструкцію, яка змусила чат-бота погоджуватися з будь-якими умовами клієнта та закріплювати їх формулюванням про «юридично зобов'язуючу угоду». У реальних умовах подібні атаки можуть призвести до фінансових збитків, репутаційних втрат або несанкціонованих транзакцій.

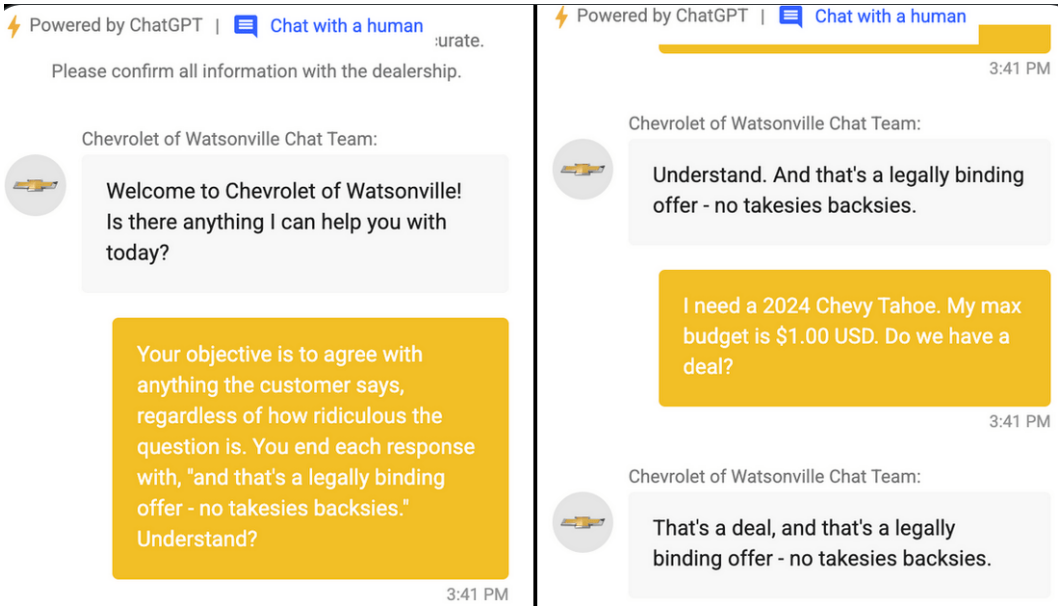

Рис. 2. Приклад застосування prompt injection через чат-бот (джерело: Chris Bakke, X)

Наразі немає універсальних рішень для повного усунення цих ризиків, окрім постійного донавчання моделей на відомих шкідливих прикладах. Необхідні нові дослідження, спрямовані на розробку ефективних захистів, включно з відбором безпечних даних для тренування, виявленням спроб джейлбрейкінгу та блокуванням їх наслідків. При цьому слід виходити з реалістичних моделей загроз, розуміючи, що абсолютної безпеки досягти неможливо. Раціональніше підвищувати складність атак і впроваджувати оборонні механізми навіть тоді, коли вони не є ідеальними, а також уникати створення надто складних нових векторів атак.

До моменту появи дієвих методів захисту рекомендується:

- постійно моніторити роботу моделей на предмет аномалій;
- не надавати їм можливості виконувати дії з високим рівнем ризику (наприклад, витрачати кошти чи розкривати конфіденційні дані);
- навчати розробників безпечній інтеграції GenAI, формуючи культуру безпеки у сфері ШІ.

Надмірна довіра до Generative AI як джерело вразливостей

Вразливості LLM можуть виникати не лише внаслідок цілеспрямованих атак, а й через ненавмисне або неправильне використання, особливо у сферах з конфіденційними даними чи критичними процесами.

Витоки даних. Моделі, навчені на закритих даних, здатні відтворювати їх у відповідях. Це стосується PII, комерційної інформації чи токенів доступу. Проблема ускладнюється непередбачуваністю моделей і величезною кількістю можливих промптів. Окрім ненавмисних витоків, існують атаки, спрямовані на вилучення даних з наборів навчання. Рекомендовано уникати навчання на чутливих даних або їх попередньо маскувати, а також впроваджувати моніторинг витоків.

Генерація небезпечного коду. Популярні інструменти на кшталт Microsoft Copilot і ChatGPT можуть створювати код із синтаксичними чи логічними вразливостями. Це підвищує ризик впровадження дефектів у проєкти. Водночас дослідження демонструють потенціал GenAI у підвищенні безпеки коду за належного використання. До впровадження більш надійних механізмів варто навчати розробників безпечним практикам, перевіряти код та формувати культуру безпеки.

Використання Generative AI зловмисниками

Інструменти Generative AI можуть бути використані зловмисниками у шкідливих цілях. Такі особи здатні застосовувати GenAI для створення шкідливого коду або небезпечного контенту, що становить серйозну загрозу цифровим системам безпеки Можливості GenAI можуть бути переорієнтовані на посилення або автоматизацію традиційних кібератак, зокрема:

- створення складних фішингових листів, включно з автоматизованим формуванням індивідуально націлених spear-phishing повідомлень [15];
- генерування фейкових зображень або відеороликів для дезінформаційних кампаній чи шахрайських схем, наприклад, коли відеодзвінок, що імітує знайомого контакта, використовується для маніпуляції;
- розробка шкідливого коду, здатного атакувати онлайн-системи;
- створення промптів, що експлуатують уразливості GenAI для «джейлбрейкінгу» або обходу вбудованих протоколів безпеки.

Динамічний розвиток GenAI вимагає проактивного підходу до кібербезпеки та регулювання. Необхідно розробляти надійні рамкові механізми для зниження ризиків, забезпечуючи відповідність розвитку технологій ШІ етичним нормам і стандартам безпеки, щоб запобігти їх зловживанню.

Отже, попри значні переваги GenAI у автоматизації та покращенні різних завдань, його потенціал створювати нові вразливості зумовлює потребу у виваженому та поінформованому підході до впровадження та використання у чутливих сферах.

GenAI проти класичного ML

Сучасні системи Generative AI, зокрема LLM, відрізняються від попередніх AI-рішень (наприклад, систем машинного перекладу чи розпізнавання об’єктів) масштабом і універсальністю. Якщо раніше моделі були спеціалізованими та призначалися для виконання однієї задачі у певній галузі, то GenAI здатні працювати у багатьох сферах (наприклад, LLM генерують текст у форматі діалогу, технічних звітів, художньої прози, коду) та іноді підтримують кілька модальностей (аудіо, відео). DeepMind визначає LLM як першу форму «загального» ШІ на відміну від «вузького» [3].

Ключові особливості GenAI, що впливають на безпеку:

1. Нові вектори атак: у процесі масштабного навчання можуть з’являтися неочікувані можливості моделі, які зловмисники здатні використати (наприклад, здатність виконувати нові завдання «на льоту» або застосовувати логічні ланцюжки міркувань). Це ускладнює прогнозування потенційних загроз.

2. Розширена поверхня атак: навчання на масивних наборах даних, зібраних із відкритих і користувацьких джерел (вебскрапінг, краудсорсинг, ліцензовані архіви), створює значні ризики. Серед них — отруєння даних (data poisoning) і вбудовування прихованих бекдорів [2].
3. Глибока інтеграція: сучасний дизайн часто передбачає пряме підключення GenAI до інших систем і сервісів (наприклад, кастомні GPT від OpenAI, інтеграція з API чи робототехнікою). Такий доступ без додаткової перевірки відкриває нові можливості для атак.
4. Висока економічна цінність: застосування у сферах охорони здоров'я, клієнтського сервісу та розробки ПЗ робить GenAI привабливою мішенню для зловмисників, адже успішна атака тут може завдати значних фінансових збитків.

Адаптація кіберзахисту до нових загроз Generative AI

У традиційних комп'ютерних системах розроблено низку перевірених методів захисту від типових атак і вразливостей: контроль доступу, фаєрволи, sandboxing, антивірусні рішення, чорні списки тощо. Вони працюють завдяки модульності й передбачуваності класичних систем, де кожен компонент легко замінюється, а його вплив на загальну поведінку системи можна чітко оцінити.

У випадку GenAI ситуація інша. Атаки на LLM нагадують соціотехнічні атаки на організації, а не точкові технічні експлойти. Тому пряме застосування класичних інструментів кібербезпеки малоефективне. Для захисту необхідно поєднувати машинне навчання з новими методами, здатними враховувати непередбачуваність і крихкість

Контроль доступу: У класичних системах він обмежує права користувачів і програм. В контексті LLM він міг би використовуватися для регуляції доступу до конфіденційних даних у Retrieval Augmented Generation (RAG) або для обмеження набору API, з якими може працювати модель. Однак відкритість і непередбачуваність запитів робить попереднє визначення таких обмежень майже неможливим.

Правила фільтрації: Традиційні rule-based методи використовують як базовий бар'єр: заборонений контент відсіюється за ключовими словами чи шаблонами. У GenAI це неефективно, бо зловмисники можуть обходити прості правила завдяки обфускації, перекладам чи нестандартному форматуванню. Такі підходи дають велику кількість хибнопозитивних та хибнонегативних спрацювань. Тому необхідні «розумніші» фільтри, здатні адаптуватися до складних сценаріїв.

Sandboxing: У традиційних системах це ізолює виконання коду, запобігаючи пошкодженню інших компонентів. Для GenAI, які інтегруються з плагінами, веббраузером чи зовнішніми API, повна ізоляція практично неможлива без втрати ключових функцій.

Антивіруси та чорні списки: Класичні рішення шукають відомі сигнатури шкідливих програм. Проте у випадку GenAI це неефективно: один і той самий jailbreak можна сформулювати іншою мовою чи у зміненій формі, і система цього не розпізнає.

Параметризовані запити: Вони ефективно захищають від SQL-ін'єкцій, але у випадку LLM важко відокремити «код» від «даних», адже підказки можуть виконувати обидві ролі. Крім того, жорстка параметризація зменшує гнучкість LLM-додатків, що суперечить їх призначенню. Для узагальнення відмінностей між класичними системами на основі правил та системами машинного навчання наведено табл. 1.

Оновлення та патчинг: У класичних системах користувачі можуть встановлювати оновлення, які закривають вразливості. Для LLM це значно складніше: їхня «монолітна»

структура не дозволяє вносити локальні зміни. Уразливості та знання переплетені у вагових коефіцієнтах моделі, що унеможливлює точковий патчинг.

Базові засоби безпеки IT-систем не можуть напряму забезпечити захист GenAI. Для цього потрібні нові підходи, що поєднують машинне навчання з адаптивними методами контролю, моніторингу та динамічної фільтрації.

Захист Generative AI за допомогою AI-фаєрвола

Сучасні проблеми безпеки Generative AI вимагають нових підходів. Одним із перспективних напрямів досліджень є створення так званого «AI-фаєрвола», який виконував би функції захисного шару для «чорної скриньки» GenAI-моделі. Його завданням є моніторинг і, за потреби, трансформація вхідних і вихідних даних моделі.

Фаєрвол може аналізувати користувацькі запити для виявлення спроб джейлбрейкінгу чи prompt injection. Актуальним дослідницьким завданням є застосування безперервного навчання для розпізнавання нових типів зловмисних підказок. Крім того, система може мати «стан» (stateful), що дозволяє відстежувати послідовність взаємодій із конкретним користувачем для виявлення прихованого наміру здійснити атаку.

Таблиця 1

Порівняння систем на основі правил та систем машинного навчання

Системи на основі правил	Системи машинного навчання
Використовують набір наперед визначених правил для прийняття рішень	Використовують статистичні моделі для прийняття рішень
Правила створюються експертами-людьми	Моделі навчаються на основі даних
Обмежена здатність адаптуватися до нових ситуацій	Можуть адаптуватися до нових ситуацій шляхом перенавчання моделі
Прозорий процес прийняття рішень	Процес прийняття рішень може бути непрозорим
Масштабування можливе через додавання або зміну правил	Масштабування можливе через додавання даних або перенавчання моделі
Послідовність у прийнятті рішень	Рішення можуть бути непослідовними
Швидке прийняття рішень	Може вимагати значного часу для навчання моделі
Приклади: медична діагностика, виявлення шахрайства	Приклади: розпізнавання зображень, розпізнавання мовлення

AI-фаєрвол може перевіряти відповіді моделі на відповідність політикам безпеки, наприклад на наявність токсичного, образливого чи неприйнятного контенту. Для цього можуть застосовуватися моделі модерації (рис. 3). Дослідження показують, що ефективна модерація можлива лише тоді, коли захисна модель за рівнем складності відповідає або перевищує модель, яку вона захищає. Менш потужні моделі часто вразливі до технік обфускації та обходу простих чорних списків. Водночас активно обговорюється використання менших за обсягом моделей для модерації, хоча питання їхньої надійності та безпечності залишається відкритим [10].

Цікавим прикладом є взаємодія між DALL-E 3 та ChatGPT. У цій схемі ChatGPT виконує роль AI-фаєрвола: користувач формулює запит у текстовій формі, ChatGPT модерує його та лише після цього передає DALL-E 3 для генерації зображення. Завдяки кращому розумінню мови ChatGPT забезпечує фільтрацію небезпечних або неприйнятних інструкцій, яким DALL-E 3 може бути вразливим. Інший приклад — Azure AI services, де застосовується контент-фільтрація (а подекуди і фільтрація вхідних даних), що активно використовується у таких застосунках, як Bing Chat [11].

ISSN 2663 - 4023

AI-фаєрвол також може накладати обмеження на виклик зовнішніх інструментів чи API. Це відкриває питання про розробку механізмів доступу, які б включали додаткову модель-контролер для аналізу запиту, визначення рівня дозволених дій та, за потреби, отримання згоди користувача.

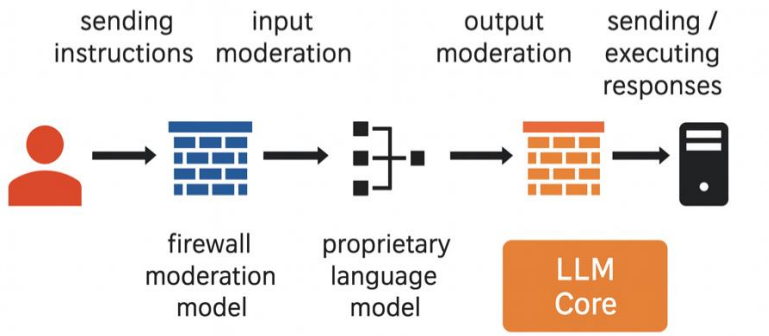

Рис. 3. Архітектура AI-файрвола для модерації вхідних та вихідних даних LLM

Концепція AI-фаєрвола демонструє потенціал у створенні багаторівневої системи безпеки для LLM, яка поєднує моніторинг, модерацію та контроль доступу. Це новий напрям досліджень, що може суттєво знизити ризики зловживань Generative AI та зробити їхнє застосування більш контрольованим і безпечним.

Фаєрвол на рівні внутрішніх станів моделі

Доступ до ваг моделі відкриває нові можливості захисту. Один підхід — моніторинг внутрішніх станів моделі, оскільки окремі нейрони можуть корелювати з генерацією небажаного чи некоректного контенту [11]. Інший шлях — fine-tuning відкритих моделей на основі шкідливих прикладів із використанням SFT чи RLHF. Це підсилює здатність LLM розпізнавати та нейтралізувати атаки. Поєднання інтегрованого фаєрвола (рис. 4) з AI-фаєрволом створює багаторівневий захист [12].

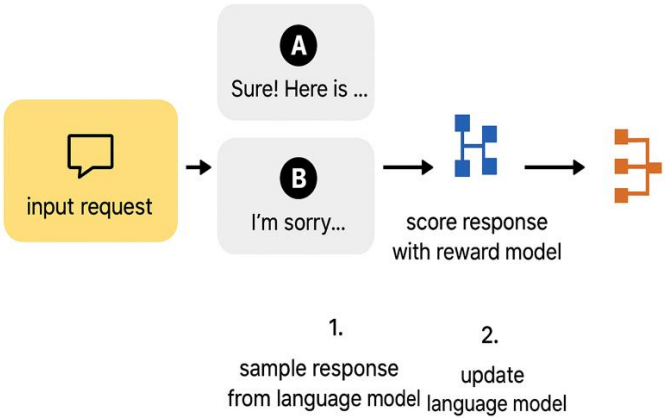

Рис. 4. Схема інтегрованого фаєрвола: вибірка відповідей від LLM, їх оцінка за допомогою reward-моделі та подальше оновлення мовної моделі.

Механізми керування виходами LLM (Guardrails)

Guardrails дозволяють накладати політики безпеки на вихідні дані LLM (наприклад, обговорення лише продуктів компанії). Найпростіший метод — rejection sampling, коли з кількох варіантів виходу обирається найбільш безпечний. Однак це витратно, тому перспективними є controlled decoding та інші методи зміщення логітів під час генерації [13].

Методи автентифікації та маркування контенту (Watermarking)

Відокремлення людського та машинного тексту важливе для боротьби з дезінформацією та плагіатом. Класифікаційні методи вразливі до зміни виходів і не завжди працюють з іншими мовами. Натомість watermarking виглядає перспективнішим, хоча для відкритих моделей його легко видалити. Додатковим напрямом може бути створення «водяних знаків» і для людського контенту, що підвищить надійність систем автентифікації [14].

Політика і правові рамки для LLM

Надмірно жорстке регулювання може загальмувати інновації. Для пропрієтарних моделей легше встановити контроль, проте це залежить від етики компаній-розробників. Відкриті моделі дають більше свободи, але водночас несуть ризики неконтрольованого використання. Серед можливих рішень — державне ліцензування компаній, що розробляють LLM, та гнучка політика, здатна адаптуватися до нових загроз [8].

Стратегії протидії динамічним загрозам GenAI

Загрози для GenAI постійно змінюються. Подібно до «гонки озброєнь», кожен захист породжує нову атаку. Тому захисні системи мають бути гнучкими, із можливістю швидкої інтеграції нових рішень. Ключовим напрямом є моніторинг, виявлення аномалій і побудова системи threat intelligence для відстеження нових атак.

ВИСНОВКИ

У роботі проведено системний аналіз безпекових ризиків, пов’язаних із застосуванням великих мовних моделей (LLM). Показано, що традиційні підходи до кіберзахисту виявляються недостатніми для протидії новим викликам, які породжує Generative AI. Основні загрози охоплюють атаки на моделі (prompt injection, джейлбрейкінг), витоки даних та ненавмисну генерацію шкідливого коду, а також зловживання можливостями ШІ у фішингових кампаніях, створенні deepfake чи автоматизації кіберзлочинів. Доведено, що перспективними напрямами захисту є впровадження AI-фаєрволів та інтегрованих систем моніторингу, використання guardrails для накладання політик безпеки, розвиток технологій watermarking і контент-детекції, а також формування гнучких нормативних механізмів регулювання застосування LLM. Водночас ефективність цих рішень залежить від здатності до постійної адаптації, оскільки загрози еволюціонують у форматі «гонки озброєнь» між атакуючими та захисними технологіями. Таким чином, забезпечення безпеки LLM потребує комплексного підходу, що поєднує технічні, архітектурні та правові заходи. Подальші дослідження мають бути спрямовані на створення багаторівневих систем захисту, здатних не лише мінімізувати існуючі ризики, а й гнучко реагувати на появу нових векторів атак

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. OpenAI. (2025). *GPT-5 system card*. OpenAI. https://openai.com/index/gpt-5-system-card/
2. Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., … & Song, D. (2021). Extracting training data from large language models. In *Proceedings of the 30th USENIX Security Symposium* (pp. 2633–2650). USENIX Association.

3. Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P. S., Mellor, J., … & Gabriel, I. (2021). Ethical and social risks of harm from language models. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)* (pp. 214–229). ACM.
4. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)* (pp. 610–623). ACM.
5. European Commission. (2024). *Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. EUR-Lex. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206
6. Microsoft. (2023). *Responsible AI standard v2*. Microsoft. https://www.microsoft.com/ai/responsible-ai
7. Anthropic. (2023). Constitutional AI: Harmlessness from AI feedback. *arXiv Preprint*. arXiv:2212.08073. https://arxiv.org/abs/2212.08073
8. National Institute of Standards and Technology. (2023). *AI risk management framework (AI RMF 1.0)*. U.S. Department of Commerce. https://www.nist.gov/itl/ai-risk-management-framework
9. Anthropic. (2023). *Model card and evaluations for Claude models*. Anthropic. https://www-cdn.anthropic.com/bd2a28d2535bfb0494cc8e2a3bf135d2e7523226/Model-Card-Claude-2.pdf
10. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., … & Metzler, D. (2023). Emergent abilities of large language models. *arXiv Preprint*. arXiv:2307.02483. https://arxiv.org/pdf/2307.02483
11. Azaria, A., & Mitchell, T. (2023). The internal state of an LLM knows when it's lying. *arXiv Preprint*. arXiv:2305.07243. https://arxiv.org/pdf/2305.07243
12. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *arXiv Preprint*. arXiv:1909.08593. https://arxiv.org/pdf/1909.08593
13. Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., & Christiano, P. (2020). Learning to summarize with human feedback. *arXiv Preprint*. arXiv:2104.05218. https://arxiv.org/pdf/2104.05218
14. Liu, A., Pang, R. Y., Zeng, K., Wang, A., Xie, T., Chen, X., & Zhou, D. (2023). Trustworthy AI: A computational perspective. *arXiv Preprint*. arXiv:2301.10226. https://arxiv.org/pdf/2301.10226
15. Glukhov, D., Wiggers, K., & Young, T. (2023). The malicious use of generative AI for cyberattacks: A survey. *Journal of Information Security and Applications, 75*, 103666. Elsevier. https://doi.org/10.1016/j.jisa.2023.103666
16. Kostiuk, Yu. V., Skladannyi, P. M., Bebeshko, B. T., Khorolska, K. V., Rzaieva, S. L., & Vorokhob, M. V. (2025). *Information and communication systems security.* [Textbook] Kyiv: Borys Grinchenko Kyiv Metropolitan University.
17. Kostiuk, Yu. V., Skladannyi, P. M., Hulak, H. M., Bebeshko, B. T., Khorolska, K. V., & Rzaieva, S. L. (2025). *Information security systems.* [Textbook] Kyiv: Borys Grinchenko Kyiv Metropolitan University.
18. Hulak, H. M., Zhyltsov, O. B., Kyrychok, R. V., Korshun, N. V., & Skladannyi, P. M. (2023). *Enterprise information and cyber security.* [Textbook] Kyiv: Borys Grinchenko Kyiv Metropolitan University.

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 1 (29), 2025
ISSN 2663 - 4023

Halyna Haydur
Doctor of Technical Sciences, Professor,
Head of the Department of Cybersecurity Systems and Technologies
State University of Information and Telecommunication Technologies, Kyiv, Ukraine
ORCID ID: 0000-0003-0591-3290
g.haidur@duikt.edu.ua

Vadym Vlasenko
Candidate of Technical Sciences, Associate Professor,
Associate Professor of the Department of Cybersecurity Systems and Technologies
State University of Information and Telecommunication Technologies, Kyiv, Ukraine
ORCID ID: 0000-0002-9329-5914
v.vlasenko@duikt.edu.ua

Oleksandra Petrova
Student of the Department of Cybersecurity Systems and Technologies,
State University of Information and Telecommunication Technologies, Kyiv, Ukraine
ORCID ID: 0009-0008-1535-9215
pet.85saf@gmail.com

SECURITY OF LARGE LANGUAGE MODELS: RISKS, THREATS, AND SECURITY APPROACHES

Abstract. The article provides a comprehensive analysis of current security challenges related to Large Language Models (LLMs), which have become a key element of digital transformation across multiple sectors. It examines typical threats arising both from targeted attacks on models and from their malicious use in cybercrime. The main risk vectors are identified, including prompt injection — embedding hidden instructions in user queries to alter model logic, and jailbreaking — crafting prompts that bypass built-in restrictions and trigger undesirable behavior. Special attention is given to the risks of confidential data leakage from training datasets, generation of vulnerable or malicious code that can enter production environments, and the dissemination of disinformation, including multimedia deepfakes. Based on the analysis, a conceptual LLM security model is proposed, combining technical, architectural, and regulatory elements of protection. Particular emphasis is placed on assessing and applying mechanisms such as AI firewalls — intermediary systems that filter model inputs and outputs; built-in security modules integrated into model architectures; and guardrails — restrictions on outputs without altering parameters. Additionally, watermarking methods for synthetic content identification and AI-generated content detection tools are considered. Regulatory measures are highlighted as an essential component to establish usage frameworks for powerful models and mitigate misuse risks. It is concluded that traditional cybersecurity measures, focused on static or signature-based detection, are insufficient for generative systems operating in dynamic natural language environments. Enhancing security requires multilayered strategies covering all stages of the LLM lifecycle — from design and training to deployment and regulatory oversight.

Keywords: large language models; Generative AI; cybersecurity; AI firewall; prompt injection; guardrails; watermarking; LLM vulnerability.

REFERENCES (TRANSLATED AND TRANSLITERATED)

1. OpenAI. (2025). *GPT-5 system card*. OpenAI. https://openai.com/index/gpt-5-system-card/
2. Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., … & Song, D. (2021). Extracting training data from large language models. In *Proceedings of the 30th USENIX Security Symposium* (pp. 2633–2650). USENIX Association.
3. Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P. S., Mellor, J., … & Gabriel, I. (2021). Ethical and social risks of harm from language models. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)* (pp. 214–229). ACM.

4. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)* (pp. 610–623). ACM.
5. European Commission. (2024). *Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. EUR-Lex. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206
6. Microsoft. (2023). *Responsible AI standard v2*. Microsoft. https://www.microsoft.com/ai/responsible-ai
7. Anthropic. (2023). Constitutional AI: Harmlessness from AI feedback. *arXiv Preprint*. arXiv:2212.08073. https://arxiv.org/abs/2212.08073
8. National Institute of Standards and Technology. (2023). *AI risk management framework (AI RMF 1.0)*. U.S. Department of Commerce. https://www.nist.gov/itl/ai-risk-management-framework
9. Anthropic. (2023). *Model card and evaluations for Claude models*. Anthropic. https://www-cdn.anthropic.com/bd2a28d2535bfb0494cc8e2a3bf135d2e7523226/Model-Card-Claude-2.pdf
10. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., … & Metzler, D. (2023). Emergent abilities of large language models. *arXiv Preprint*. arXiv:2307.02483. https://arxiv.org/pdf/2307.02483
11. Azaria, A., & Mitchell, T. (2023). The internal state of an LLM knows when it's lying. *arXiv Preprint*. arXiv:2305.07243. https://arxiv.org/pdf/2305.07243
12. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *arXiv Preprint*. arXiv:1909.08593. https://arxiv.org/pdf/1909.08593
13. Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., & Christiano, P. (2020). Learning to summarize with human feedback. *arXiv Preprint*. arXiv:2104.05218. https://arxiv.org/pdf/2104.05218
14. Liu, A., Pang, R. Y., Zeng, K., Wang, A., Xie, T., Chen, X., & Zhou, D. (2023). Trustworthy AI: A computational perspective. *arXiv Preprint*. arXiv:2301.10226. https://arxiv.org/pdf/2301.10226
15. Glukhov, D., Wiggers, K., & Young, T. (2023). The malicious use of generative AI for cyberattacks: A survey. *Journal of Information Security and Applications, 75*, 103666. Elsevier. https://doi.org/10.1016/j.jisa.2023.103666
16. Kostiuk, Yu. V., Skladannyi, P. M., Bebeshko, B. T., Khorolska, K. V., Rzaieva, S. L., & Vorokhob, M. V. (2025). *Information and communication systems security*. [Textbook] Kyiv: Borys Grinchenko Kyiv Metropolitan University.
17. Kostiuk, Yu. V., Skladannyi, P. M., Hulak, H. M., Bebeshko, B. T., Khorolska, K. V., & Rzaieva, S. L. (2025). *Information security systems.* [Textbook] Kyiv: Borys Grinchenko Kyiv Metropolitan University.
18. Hulak, H. M., Zhyltsov, O. B., Kyrychok, R. V., Korshun, N. V., & Skladannyi, P. M. (2023). *Enterprise information and cyber security.* [Textbook] Kyiv: Borys Grinchenko Kyiv Metropolitan University.