Міністерство освіти і науки України
Київський національний університет імені Тараса Шевченка

Факультет інформаційних технологій
Кафедра кібербезпеки та захисту інформації

ДОПУСТИТИ ДО ЗАХИСТУ:
В.о. завідувача кафедри
кібербезпеки та захисту інформації
_____ Іван ПАРХОМЕНКО
«__» червня 2025 р.

ПОЯСНЮВАЛЬНА ЗАПИСКА
кваліфікаційної роботи

галузь знань 12 Інформаційні технології
(шифр і назва галузі знань)

спеціальність 125 Кібербезпека
(код і назва спеціальності)

освітній ступень бакалавр

освітня програма Кібербезпека
(назва освітньо-професійної програми)

на тему: «Механізм оцінювання ризиків використання моделей штучного інтелекту в зловмисних намірах»

Виконавець: студент IV курсу, групи КБ-42

_____ **Артем КИРЛИЦЯ**
(підпис) (ім'я, прізвище)

	Підпис	**Ім'я, прізвище**
Керівник		Юрій ЩЕБЛАНІН
Нормоконтроль		Яніна ШЕСТАК

Київ 2025

Міністерство освіти і науки України
Київський національний університет імені Тараса Шевченка

Факультет інформаційних технологій
Кафедра кібербезпеки та захисту інформації

ЗАТВЕРДЖЕНО:
В.о. завідувача кафедри
кібербезпеки та захисту інформації
_____ Іван ПАРХОМЕНКО
«29» листопада 2024 р.

ЗАВДАННЯ
на виконання кваліфікаційної роботи

спеціальності 125 Кібербезпека
(код і назва спеціальності)

освітня програма Кібербезпека
(назва освітньо-професійної програми)

Студенту **КБ-42** **Кирлиці Артему Олександровичу**
(група) (прізвище ім'я по батькові)

Тема кваліфікаційної роботи Механізм оцінювання ризиків використання моделей штучного інтелекту в зловмисних намірах

1. ПІДСТАВИ ДЛЯ ПРОВЕДЕННЯ РОБОТИ

Тема кваліфікаційної роботи затверджена на засіданні кафедри кібербезпеки та захисту інформації протокол №6 від 28.11.2024 р.

2. ВИХІДНІ ДАНІ ДЛЯ ПРОВЕДЕННЯ РОБІТ

Комплекс технічних, організаційних і правових рекомендацій

3. ЗМІСТ РОЗРАХУНКОВО-ПОЯСНЮВАЛЬНОЇ ЗАПИСКИ

Необхідно ознайомитись з принципами роботи LLM та мультимодальних генераторів, проаналізувати сценарії їх зловмисного використання та пов'язані ризики, розробити комплекс контрзаходів.

4. ВИМОГИ ДО РЕЗУЛЬТАТІВ ВИКОНАННЯ РОБОТИ

Практична цінність Сформовано рекомендації та механізм оцінювання ризиків зловмисного використання ШІ, що можуть бути впроваджені у політики кібербезпеки організацій.

5. ДАТА ВИДАЧІ ЗАВДАННЯ

Дата видачі завдання: 29 листопада 2024 року

Завдання видав _____ (підпис) Юрій ЩЕБЛАНІН (ім'я, прізвище)

Завдання прийняв до виконання _____ (підпис) Артем КИРЛИЦЯ (ім'я, прізвище)

КАЛЕНДАРНИЙ ПЛАН

№ п/п	Найменування етапів робіт	Строки виконання робіт (початок-кінець)	Відмітка про виконання
1	Уточнення постановки задачі	29.11.2024 – 03.12.2024	*виконано*
2	Аналіз літератури	04.12.2024 – 05.02.2025	*виконано*
3	Вивчення сучасних моделей штучного інтелекту та векторів атак	06.02.2025 – 15.02.2025	*виконано*
4	Систематизація типових ризиків зловмисного використання ШІ	16.04.2025 – 23.05.2025	*виконано*
5	Розробка структури оцінювання та класифікації загроз	24.04.2025 – 02.05.2025	*виконано*
7	Розробка практичних рекомендацій і концептуальної моделі оцінювання ризиків	12.05.2020 – 25.05.2021	*виконано*
8	Формалізація та опис власної пропозиції	25.05.2025 – 28.05.2025	*виконано*
9	Оформлення пояснювальної записки	28.05.2025 – 13.06.2025	*виконано*

Завдання видав _____ (підпис) Юрій ЩЕБЛАНІН (ім'я, прізвище)

Завдання прийняв до виконання _____ (підпис) Артем КИРЛИЦЯ (ім'я, прізвище)

Термін подання кваліфікаційної роботи до ЕК 13 червня 2025 року

РЕФЕРАТ

Кваліфікаційна робота складається зі вступу, трьох розділів, загальних висновків, списку використаних джерел, додатків, має 80 сторінок основного тексту, 8 таблиць та 6 рисунків. Список використаних джерел містить 48 найменувань і займає 5 сторінок.

Метою роботи є покращення кіберзахисту від загроз зловмисного використання моделей штучного інтелекту.

Для досягнення зазначеної мети у роботі були поставлено наступні завдання:

- проаналізувати ризики, що виникають унаслідок зловмисного безпосереднього використання генеративних моделей штучного інтелекту.
- дослідити механізми, якими ШІ підсилює складні кіберзагрози .
- розробити рекомендації щодо впровадження технічних, організаційних та нормативно-правових механізмів протидії зазначеним ризикам.

Об'єктом дослідження є процес зловмисної експлуатації генеративних моделей штучного інтелекту у кіберпросторі.

Предметом дослідження є методи, алгоритми та організаційно-правові механізми оцінювання й мінімізації ризиків, що виникають унаслідок зловмисного використання генеративних моделей штучного інтелекту у кіберпросторі.

Практичною цінністю отриманих результатів є можливість використання напрацьованих методик для вдосконалення політик інформаційної безпеки та засобів захисту від нових кіберзагроз.

Ключові слова: штучний інтелект, великі мовні моделі, генеративні системи, кіберзагрози, deepfake, соціальна інженерія, дезінформація, фішинг, корпоративна безпека, репутаційні ризики.

ЗМІСТ

ВСТУП……..8
РОЗДІЛ 1 ТЕОРЕТИЧНІ ОСНОВИ ФУНКЦІОНУВАННЯ СУЧАСНИХ МОДЕЛЕЙ ШТУЧНОГО ІНТЕЛЕКТУ 9
1.1 Основні напрями розвитку машинного навчання 9
1.2 Огляд поточного стану ринку штучного інтелекту 14
1.2.1 OpenAI та його інструментарій 14
1.2.2 Microsoft Copilot як альтернативний розвиток ChatGPT 17
1.2.3 Google Gemini: Архітектура, функціональність та застосування 18
1.3 Генеративні моделі ШІ: принципи дії та сфери застосування 20
1.3.1 Розгляд структурно-функціональних підходів до реалізації генеративного ШІ, зокрема LLM-моделей як GPT, PaLM, Claude та ін. . 20
1.3.2 Принципи функціонування генеративних моделей у контексті обробки різнорідних даних: тексту, зображень, відео та аудіосигналів .. 24
1.3.3 Deepfake-технологія як приклад синтетичного мультимедійного контенту 26
Висновки за розділом 1 28
РОЗДІЛ 2 АНАЛІЗ РИЗИКІВ, ПОВ’ЯЗАНИХ ІЗ ЗЛОВМИСНИМ ВИКОРИСТАННЯМ ШІ 30
2.1 Кіберзагрози, що реалізуються за допомогою ШІ 30
2.1.1 Автоматизовані кібератаки 30
2.1.2 Генерація фішингових повідомлень і листів. 34
2.2 Деструктивне застосування дипфейків як похідної generative ШІ 35
2.2.1 Шантаж і вимагання 36
2.2.2 Фальсифікація відео- та аудіодоказів 36
2.2.3 Вплив на суспільно-політичні процеси 40
2.3 Інформоперації за допомогою ШІ у соціальних мережах 43
2.4 Ризики для корпоративного та державного сектору 44
2.4.1 Компрометація внутрішніх комунікацій та витоки даних 45

2.4.2 Шахрайські фіноперації 47
2.4.3 Репутаційні збитки та зниження довіри клієнтів 50
Висновки за розділом 2 52
РОЗДІЛ 3 СИСТЕМНИЙ ПІДХІД ДО ВИЯВЛЕННЯ ТА ПРОТИДІЇ ЗАГРОЗАМ ШІ В КІБЕРПРОСТОРІ 54
3.1 Технічні методи виявлення фальшивого контенту 54
3.2 Рекомендації зі створення ПЗ для захисту на державному рівні 56
3.3 Організаційно-процедурні заходи з протидії ризиків ШІ 63
3.4 Регуляторно-правові механізми 66
3.5 Міжнародне співробітництво у сфері нейтралізації загроз ШІ 67
ВИСНОВКИ 72
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ 74
ДОДАТОК А 80
ДОДАТОК Б 82

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ ТА СКОРОЧЕННЬ

ШІ	−	штучний інтелект
КСЗШІ	−	комплексна система захисту від ші
AI	−	Artificial Intelligence − штучний інтелект
LLM	−	Large Language Model − велика мовна модель;
GAN	−	Generative Adversarial Network − генеративно-змагальна мережа
GPT	−	Generative Pre-trained Transformer − генеративний попередньо навчений трансформер
RAG	−	Retrieval-Augmented Generation − Генерація з доповненим пошуком

ВСТУП

Стрімка цифрова трансформація суспільства зумовила глибоку інтеграцію штучного інтелекту в бізнес-процеси, державне управління та повсякденну комунікацію, що підтверджується дослідженнями Stanford HAI, MIT Media Lab та звітами ENISA. Найбільш динамічно розвиваються генеративні моделі — великі мовні моделі (GPT-4, PaLM 2, Gemini), дифузійні мережі й deepfake-технології на базі GAN, здатні автономно продукувати текстовий, візуальний і аудіоконтент.

Водночас література останніх років (Goodfellow, Chollet, Brundage) фіксує зростання інцидентів зловмисного використання таких моделей: від високоточних фішингових кампаній до масштабних інформаційних операцій із застосуванням дипфейків, що підривають цілісність даних, приватність і довіру до цифрових сервісів.

Актуальність обраної теми полягає у потребі розробити комплексний механізм виявлення й нейтралізації ШІ-індукованих кіберзагроз.

Метою роботи є створення багаторівневого підходу до оцінювання ризиків зловмисного застосування генеративних моделей штучного інтелекту. Для її досягнення передбачено аналіз сучасних загроз, оцінку їхнього впливу на CIA-трикутник, обґрунтування методів детекції синтетичного контенту та формулювання технічних, організаційних і правових контрзаходів.

Об’єктом дослідження є процес зловмисної експлуатації генеративних моделей ШІ у кіберпросторі, предметом — методи і алгоритми оцінювання та мінімізації пов’язаних ризиків; у роботі застосовано статистичне моделювання, One-Class SVM, глибокі CNN-детектори й експертне опитування.

Практичне значення результатів полягає у можливості інтеграції розробленого механізму в SOC-платформи й державно-корпоративні політики кібербезпеки.

РОЗДІЛ 1
ТЕОРЕТИЧНІ ОСНОВИ ФУНКЦІОНУВАННЯ СУЧАСНИХ МОДЕЛЕЙ ШТУЧНОГО ІНТЕЛЕКТУ

1.1 Основні напрями розвитку машинного навчання

Обидва види моделей ШІ (тобто LLM моделі та DeepFake моделі), що будуть розглянуті в даній дипломній роботі, є різними типами нейромереж. Нейромережі , в свою чергу, створюються за рахунок машинного навчання (ML від Machine Learning). Даний термін був запропонований у 1959 році Артуром Семюелем, співробітником IBM та піонером у галузі комп'ютерних ігор та штучного інтелекту. Синонім «самонавчальні комп'ютери» також використовувався в цей період [1].

Машинне навчання— це підгалузь штучного інтелекту, яка надає комп'ютерам здатність самостійно навчатися на основі наданих їм даних без явного програмування. Цей підхід дозволяє системам адаптуватися до нових вхідних даних і виконувати завдання з високою точністю. З моменту свого виникнення машинне навчання пройшло значний шлях розвитку, охоплюючи різноманітні методи та алгоритми. Станом на сьогодні, основні напрями МН являють собою супервізоване, несупервізоване навчання, навчання з підкріпленням.

Основна різниця машинного навчання від традиційного в тому, що у традиційному програмуванні програміст вручну надає комп'ютеру конкретні інструкції на основі свого розуміння та аналізу проблеми. Якщо дані або проблема змінюються, програмісту потрібно вручну оновити код. На відміну від цього, у машинному навчанні процес автоматизований: ми передаємо дані комп'ютеру, і він пропонує рішення (тобто модель) без чітких інструкцій щодо того, як це зробити. Оскільки модель машинного навчання навчається сама, вона може обробляти нові дані або нові сценарії.

Існує тісний зв'язок між машинним навчанням та стисненням даних. Система, яка прогнозує апостеріорні ймовірності послідовності, враховуючи всю її історію, може бути використана для оптимального стиснення даних (за допомогою арифметичного кодування на розподілі вихідних даних). І навпаки, оптимальний компресор може бути використаний для прогнозування (шляхом знаходження символу, який найкраще стискається, враховуючи попередню історію). Ця еквівалентність була використана як обґрунтування для використання стиснення даних як еталону для "загального інтелекту". Прикладами програмного забезпечення для стиснення аудіо/відео на базі штучного інтелекту є NVIDIA Maxine, AIVC.

Прикладами програмного забезпечення, яке може виконувати стиснення зображень на базі штучного інтелекту, є OpenCV, TensorFlow, Image Processing Toolbox (IPT) від MATLAB та High-Fidelity Generative Image Compression [2].

У несупервізованому машинному навчанні кластеризація може бути використана для стиснення даних шляхом групування подібних точок даних у кластери. Цей метод спрощує обробку великих наборів даних, яким бракує попередньо визначених міток, і знаходить широке застосування в таких галузях, як стиснення зображень [3].

Моделі великих мов (LLM) також є ефективними компресорами даних без втрат для деяких наборів даних, як продемонструвало дослідження DeepMind з моделлю Chinchilla 70B. Розроблений DeepMind, Chinchilla 70B ефективно стискав дані, перевершуючи традиційні методи, такі як Portable Network Graphics (PNG) для зображень та Free Lossless Audio Codec (FLAC) для аудіо. Він досяг стиснення зображень та аудіоданих до 43,4% та 16,4% від їх початкового розміру відповідно. Однак є певні підстави для занепокоєння тим, що набір даних, який використовується для тестування, перетинається з навчальним набором даних LLM, що робить можливим, що модель Chinchilla 70B є ефективним інструментом стиснення лише для даних, на яких вона вже була навчена.

Крім цих основних напрямів, сучасне машинне навчання включає також глибоке навчання (Deep Learning), яке використовує багатошарові нейронні

мережі для обробки складних даних, таких як зображення, мова та текст. Інші важливі напрями включають федеративне навчання, яке дозволяє навчати моделі на розподілених даних без їх централізації, що є важливим для збереження конфіденційності.

Супервізоване навчання

Супервізоване навчання, від англійського supervised learning, є однією з фундаментальних парадигм машинного навчання, що передбачає побудову функціональної залежності між вхідними даними та відповідними вихідними мітками на основі навчального набору даних. Цей підхід ґрунтується на припущенні, що існує певна закономірність, яку можна виявити шляхом аналізу пар "вхід–вихід", де кожному вхідному вектору відповідає відома цільова змінна.

Процес супервізованого навчання включає кілька ключових етапів: формування навчального набору даних, вибір відповідного представлення ознак, визначення структури моделі та алгоритму навчання, а також оцінювання продуктивності моделі на тестовому наборі даних. Зазвичай, навчальний набір складається з великої кількості прикладів, де кожен приклад містить вхідні дані та відповідну мітку. Метою є побудова моделі, яка здатна узагальнювати знання з навчального набору та робити точні передбачення на нових, невідомих даних. Супервізоване навчання поділяється на два основні типи задач: класифікацію та регресію. Класифікація передбачає передбачення дискретних міток, наприклад, визначення, чи є електронний лист спамом. Регресія ж спрямована на передбачення неперервних значень, таких як прогнозування ціни нерухомості на основі її характеристик [4].

У процесі розробки великих мовних моделей (LLM), таких як ChatGPT, Gemini та інші, супервізоване навчання відіграє ключову роль на етапі тонкого налаштування (fine-tuning). Після початкового передтренування на великих обсягах неанотованих текстових даних, моделі проходять стадію супервізованого навчання, де використовуються пари "запит-відповідь", створені людьми. Цей етап дозволяє моделі адаптуватися до конкретних завдань, таких як ведення діалогу, відповіді на запитання або генерація тексту в певному

стилі. Зокрема, у випадку ChatGPT, людські тренери моделюють як користувача, так і асистента, формуючи приклади діалогів, які потім використовуються для навчання моделі. Це сприяє формуванню більш релевантних та контекстуально точних відповідей [5].

Спочатку до алгоритму задаються необроблені дані, які обробляються ним під контролем супервайзора (спостерігача) і, за необхідності, введення додаткових уточнювальних даних**.**

У сфері виявлення та протидії deepfake-технологіям супервізоване навчання також знаходить широке застосування. Моделі навчаються розпізнавати характерні ознаки підроблених зображень або відео на основі міток, що вказують на автентичність або штучність контенту. Це дозволяє створювати ефективні системи для автоматичного виявлення фальсифікацій та забезпечення інформаційної безпеки.

Попри свою ефективність, супервізоване навчання має певні обмеження, як наприклад процес збору та анотації великих обсягів даних є трудомістким і потребує значних ресурсів. Крім того, моделі можуть бути схильні до перенавчання, коли вони надто точно відображають навчальні дані, але втрачають здатність до узагальнення на нових прикладах. Це стимулює розвиток альтернативних підходів, таких як напівсупервізоване та самонавчання, які поєднують переваги супервізованого та несупервізованого навчання для досягнення кращої продуктивності. Також на Рисунок Б-1 додатку Б було наведено графічне відображення процесу супервізованого навчання.

Несупервізоване навчання

Як випливає з назви, несупервізоване навчання, або ж «навчання без учителя» використовує алгоритми самонавчання — вони навчаються без будь-яких міток чи попереднього навчання. Натомість моделі надаються необроблені, немарковані дані, і вона повинна виводити власні правила та структурувати інформацію на основі подібностей, відмінностей та закономірностей без чітких інструкцій щодо роботи з кожним елементом даних.

Одним із ключових застосувань несупервізованого навчання є кластеризація, яка дозволяє групувати дані за схожими характеристиками. Наприклад, при аналізі метеорологічних даних алгоритми кластеризації, такі як K-середніх, можуть автоматично виявляти типові погодні умови, групуючи дні за подібними параметрами, такими як температура, вологість та швидкість вітру.

У контексті розробки великих мовних моделей (LLM), таких як ChatGPT та Gemini, несупервізоване навчання відіграє вирішальну роль на етапі попереднього тренування. Під час цього етапу моделі обробляють великі обсяги неанотованих текстових даних, навчаючись передбачати наступне слово в реченні на основі контексту. Цей процес дозволяє моделям формувати глибокі внутрішні представлення мови, що забезпечує їх здатність до генерації зв'язного та контекстуально релевантного тексту. Також у сфері виявлення та протидії технологіям deepfake несупервізоване навчання також знаходить широке застосування. Зокрема, дозволяють моделям виявляти невідповідності між аудіо- та відеопотоками, що є характерними ознаками фальсифікованого контенту. Ці підходи не потребують великих обсягів анотованих даних, що значно спрощує процес навчання моделей для виявлення deepfake. Також на Рисунок Б-2 додатку Б було наведено графічне відображення процесу несупервізованого навчання [6].

Навчання з підкріпленням

На відміну від контрольованого та неконтрольованого (тобто супервізованого чи несупервізованого), навчань де моделі навчаються на основі міток або структури даних, RL ґрунтується на взаємодії агента з середовищем, де агент отримує зворотний зв'язок у вигляді винагороди або штрафу за свої дії. Мета навчання з підкріпленням полягає в тому, щоб агент вивчив оптимальну (або майже оптимальну) політику, яка максимізує функцію винагороди або інший наданий користувачем сигнал підкріплення, що накопичується з негайних винагород. Це схоже на процеси, що відбуваються в психології тварин - наприклад, біологічний мозок запрограмований інтерпретувати такі сигнали, як біль і голод, як негативні підкріплення, а задоволення та споживання їжі – як

позитивні. За деяких обставин тварини навчаються переймати поведінку, яка оптимізує ці винагороди.

У контексті RL агент сприймає стан середовища, виконує певну дію та отримує відповідну винагороду, після чого оновлює свою стратегію дій з метою максимізації сукупної винагороди в довгостроковій перспективі.

Існують два основних підходи до реалізації алгоритмів навчання з підкріпленням: моделюючі (model-based) та безмодельні (model-free) [7].

Безмодельні методи (Model-Free Reinforcement Learning) дозволяють агенту навчатися без явної моделі середовища, спираючись лише на досвід взаємодії.

Моделюючі методи (Model-Based Reinforcement Learning) передбачають побудову моделі середовища, яка описує ймовірності переходів між станами та функцію винагороди. Для кращого сприйняття було наведено графічне відображення архітектури типів машинного навчання на Рисунок Б-3 Додатку Б.

1.2 Огляд поточного стану ринку штучного інтелекту

Після огляду моделей навчання є змістовним розглянути безпосередньо архітектури моделей ШІ від провідних компаній, що доступні станом на сьогодні .

1.2.1 OpenAI та його інструментарій

OpenAI — це дослідницька структура, що спеціалізується на створенні систем штучного інтелекту загального призначення, орієнтованих на безпечну та етичну взаємодію з людиною. Основною метою компанії є забезпечення того, щоб розвиток штучного загального інтелекту слугував на благо всього людства. Її діяльність охоплює фундаментальні дослідження у сфері ШІ, розробку безпечних технологічних рішень та поширення отриманих результатів із дотриманням принципів відкритості та суспільної користі.

ChatGPT, розроблений компанією OpenAI, є одним із найвідоміших прикладів застосування великих мовних моделей (Large Language Models, LLMs) у сфері штучного інтелекту. Цей інструмент базується на архітектурі Generative Pre-trained Transformer (GPT), яка використовує механізм самоуваги (self-attention) для обробки та генерації тексту.

В червні 2018 року була представлена GPT-1 - перша ітерація серії GPT і складалася зі 117 мільйонів параметрів, саме вона заклало фундаментальну архітектуру для ChatGPT, як ми її знаємо сьогодні.

Архітектурно ChatGPT реалізований як глибока нейронна мережа, що складається з багатьох шарів трансформерів. Це дозволяє моделі обробляти контекстні залежності в тексті та генерувати зв'язні та релевантні відповіді.

Базові функціональні можливості ChatGPT включають генерацію тексту, переклад, резюмування, відповідь на запитання, написання коду та інші завдання, що потребують розуміння природної мови. Модель також може бути адаптована до специфічних завдань шляхом донавчання на спеціалізованих датасетах, що розширює її застосування в різних галузях, таких як освіта, охорона здоров'я, юриспруденція та інші.

Нижче наведено коротку характеристику актуальних версій моделей ChatGPT, актуальних на початок 2025 року:

- GPT-3.5 Turbo — базова модель з акцентом на швидкодію і низьку вартість. Працює з контекстним вікном у 16 385 токенів, але базується на знаннях до вересня 2021 року.
- GPT-4 Turbo — вдосконалена версія з підтримкою мультимодальних введень (текст + зображення), розширеним контекстом (128 000 токенів) і більш глибокими логічними зв’язками. Навчена на даних до грудня 2023 року, демонструє високий рівень достовірності відповіді та зменшену схильність до "галюцинацій".
- GPT-4o / GPT-4o mini — новітні моделі з покращеною швидкодією, контекстом до 128 000 токенів та глибшими когнітивними функціями. Вони

зберігають баланс між якістю, вартістю (GPT-4o mini дешевший) та мультимодальністю.

- OpenAI o1-preview / o1-mini — моделі, оптимізовані для STEM-навантаження з глибокою логічною обробкою, особливо в математиці, програмуванні та науках. Характеризуються високим рівнем міркування, але мають вищу затримку відповіді через триваліший процес "обдумування".
- o3 Series (2025) — останнє покоління моделей з акцентом на продуктивність і точність у наукових задачах. Вони забезпечують значно швидшу обробку, порівняно з попередниками, маючи контекст 128 000 токенів та розширені ліміти на вихідні дані.
- GPT-4.5 (Orion) — перехідна модель до GPT-5, яка поєднує переваги GPT-4 Turbo та o-серії. Вона зберігає швидкість відповіді при збільшеній точності і глибшому розумінні складних запитів. Навчена на даних до січня 2025 року.
- GPT-5 — флагманське рішення 2025 року, яке уніфікує GPT- та o-серії в одну архітектуру з можливістю налаштування рівня інтелекту відповідно до підписки. Характеризується високою адаптивністю, логічністю та підтримкою складних мультимодальних задач. [8]

Sora AI як відокремлення від ChatGPT

Sora — це генеративна модель штучного інтелекту, розроблена компанією OpenAI, яка здатна створювати відео високої якості на основі текстових описів.

Вперше модель була представлена у лютому 2024 року та одразу привернула увагу своєю здатністю генерувати реалістичні відео тривалістю до 60 секунд на основі текстових запитів користувачів. Sora базується на поєднанні дифузійних моделей (diffusion models) та трансформерів, що дозволяє їй ефективно обробляти та генерувати відео з високою точністю. Ключовим елементом є використання "просторово-часових патчів" (spacetime patches), які дозволяють моделі аналізувати та відтворювати динаміку руху та взаємодію об'єктів у відео.

Процес навчання Sora включав обробку великої кількості відео та зображень, що дозволяє моделі навчитися розуміти фізичні властивості світу, такі як гравітація, освітлення та взаємодія об'єктів. Для цього використовується попередньо навчена варіаційна автокодерна мережа (VAE) для зменшення розмірності вхідних даних та трансформерна архітектура для обробки латентних представлень [9].

Ключовим елементом Sora є використання "просторово-часових патчів" (spacetime patches), які функціонують як токени, аналогічно до слів у мовних моделях. Цей підхід базується на дослідженнях Google DeepMind щодо Vision Transformers (ViT) та їх розширення NaViT. Замість традиційного підходу, де зображення розбиваються на фіксовані патчі, Sora адаптує розмір патчів відповідно до вхідних даних, що дозволяє моделі ефективно обробляти відео з різними роздільними здатностями та аспектними співвідношеннями [10].

1.2.2 Microsoft Copilot як альтернативний розвиток ChatGPT

Після презентації компанією OpenAI GPT-3 у 2019 році, що стала новою ступінню еволюції нейромереж та мовних моделей, багато провідних IT-компаній зацікавились можливістю придбання та інвестицій у OpenAI та ChatGPT. Одним з найбільших інвесторів стали Microsoft, вклавши 10 мільярдів доларів США у компанію OpenAI, а також заключили договір про партнерство і співпрацю з OpenAI. Саме це і дало компанії Microsoft можливість початку розробки Microsoft Copilot, що в майбутньому замінить Microsoft Bing Chat [11].

Microsoft Copilot базується на моделі Prometheus, яка поєднує можливості GPT-4 від OpenAI з пошуковими можливостями Bing. Ця модель використовує компонент Orchestrator, що ітеративно генерує пошукові запити, об'єднуючи дані з Bing з розумінням і генеративними можливостями GPT-моделей.

У технічному аспекті вона спирається на використання великих LLM, що хостяться через Azure OpenAI Service, а також інтегрується із внутрішніми джерелами даних за допомогою Microsoft Graph API — платформи, яка агрегує

інформацію з Microsoft 365 (електронна пошта, календарі, документи, чати тощо). Однією з ключових інноваційних складових архітектури Copilot є використання семантичного індексу, що реалізує пошук і співставлення запитів користувача на основі векторних подань даних.

Ключовим елементом є Microsoft Graph — API, який забезпечує доступ до організаційних даних, таких як електронна пошта, документи, календарі та інше [12].

Значущим кіберінцидентом є проєкт Microsoft Recall, як інтеграція ШІ Copilot із всією ОС Windows, яка мала збирати дані з абсолютно всього, що робив користувач на ПК, безперервно створюючи скріншоти (знімки екрану) його дій, створюючи власну базу даних для подальшого пошуку цих дій в історії користувача. Окрім очевидних критичних проблем безпеки в самому задумі, проблема також полягала в тому, що ці бази даних не були зашифрованими надійним шифруванням окрім як BitLocker [13].

1.2.3 Google Gemini: Архітектура, функціональність та застосування

Google Gemini є флагманською мультимодальною моделлю штучного інтелекту нового покоління, розробленою у рамках спільних зусиль підрозділів Google DeepMind та Google Research. Також варто зазначити, що ця версія на початку публічно називалась Bard, і була перейменована лише після злиття з іншим ШІ-проєктом Google, Duet AI. Архітектурно ця модель є результатом масштабної науково-інженерної ініціативи, спрямованої на побудову уніфікованої системи, здатної інтегрувати різнорідні типи даних, зокрема текст, зображення, аудіо, відео та програмний код, у єдине обчислювальне середовище.

У порівнянні з попередніми підходами, які передбачали поєднання спеціалізованих компонентів для окремих модальностей, Gemini реалізує нативну мультимодальність. Це означає, що вже на етапі попереднього навчання модель обробляє різні типи інформації одночасно, внаслідок чого досягається глибше семантичне узгодження між модальностями. Таке рішення значно

підвищує ефективність генерації висновків у складних міждисциплінарних сценаріях, включаючи наукові дослідження, фінансовий аналіз і обробку великих обсягів неструктурованих даних [14].

Початкова реалізація моделі Gemini 1.0 включає три масштабні конфігурації: Ultra, Pro та Nano. Gemini Ultra призначена для виконання високоскладних когнітивних завдань з поглибленим логічним та абстрактним мисленням. Версія Pro орієнтована на масштабоване застосування у прикладних сервісах, зокрема в корпоративних IT-інфраструктурах. Модель Gemini Nano оптимізована для роботи на пристроях із обмеженими обчислювальними ресурсами, зокрема на смартфонах, де вона використовується в системах на кшталт Android AICore [15].

З технічної точки зору модель базується на архітектурі трансформера з дифузійними механізмами генерації, що дозволяє ефективно працювати з даними у високій роздільній здатності та великій кількості часових і просторових патернів.

Функціональні можливості моделі на сьогодення охоплюють генерацію тексту, кодування та декодування аудіоінформації, переклад, обробку зображень і відео, а також інтерпретацію багатомодальних запитів. Запроваджені у версії 1.5 функції, як-от Gemini Live, Gems (кастомізація моделей під специфічні запити), розширення API Google Apps та функціонал JSON-виводу, свідчать про наближення цих моделей до універсальних цифрових агентів, здатних до виконання складних когнітивних завдань у режимі реального часу [16].

Інноваційний генератор мультимедіа від Google – VEO3

У травні 2025 було представлено нову генеративну text-to-video модель від Google – VEO3.

Veo 3 значно перевершує попередні відеогенератори. Так, якщо дослідницька модель Google Lumiere (2024) використовувала просторово-часову U-Net-архітектуру для створення кадрів у реальному часі, забезпечуючи узгоджений рух, але не мала вбудованого звуку, то Veo 3 успадковує цю якість та додає синтез аудіо і глибокий контроль над кожним елементом сцени.

У порівнянні з аналогами від інших компаній, Veo 3 має значні переваги: OpenAI Sora (2024) генерує реалістичні відео до 1080p тривалістю приблизно ~20 секунд, але без власного звукового супроводу, тоді як Veo 3 об'єднує високу роздільну здатність, фізичну реалістичність рухів та мультимедійну інтеграцію, що робить її передовим інструментом для творчого відеомонтажу, але водночас і потенційно небезпечним у руках зловмисників.

Модель забезпечує фотореалістичні відео високої чіткості (до 4K). Вихідні кадри демонструють плавний рух, природне освітлення і правильні фізичні властивості сцени (наприклад, моделі людей з п'ятьма пальцями на руках). За заявами DeepMind, Veo 3 «поставляє найвищу якість» і відмінно дотримується реалістичності й заданого сюжет [17].

1.3 Генеративні моделі ШІ: принципи дії та сфери застосування

Генеративні моделі штучного інтелекту сьогодні є ключовим напрямом розвитку сучасних інформаційних технологій, оскільки вони забезпечують здатність до створення, трансформації та адаптації нових даних на основі складних статистичних закономірностей. Основою більшості таких моделей є архітектури глибокого навчання, які дозволяють відтворювати структуру та семантику реальних даних, що зумовлює їхню універсальність у численних сферах — від генерації природної мови та зображень до синтезу звуку та відео.

У цьому підрозділі розглядаються принципи функціонування генеративних моделей, їхня роль у вирішенні широкого спектра задач, а також їх функціональні можливості, що тягнуть за собою ризики.

1.3.1 Розгляд структурно-функціональних підходів до реалізації генеративного ШІ, зокрема LLM-моделей як GPT, PaLM, Claude та ін.

Великі мовні моделі (Large Language Models, LLM) являють собою тип нейронних мереж, здатних прогнозувати ймовірність появи наступного токена в

послідовності, виходячи з контексту попередніх елементів. Вони виявляють емерджентні властивості, що дає змогу вирішувати широкий спектр складних мовних завдань — від генерації тексту до перекладу, пошуку відповідей і логічного висновування. Основу їхньої архітектури становить механізм самоуваги, який забезпечує здатність до обробки довгих послідовностей та виявлення віддалених семантичних зв’язків у тексті.

Архітектура Transformer (трансформерів) стала фундаментальною основою для більшості сучасних великих мовних моделей (LLM), завдяки своїй здатності забезпечувати паралельну обробку інформації та ефективне моделювання довготривалих залежностей у текстових послідовностях. Центральний механізм самоуваги (self-attention) дозволяє кожному елементу послідовності (токену) взаємодіяти з усіма іншими елементами, формуючи комплексні контекстуальні представлення. Багатоголовий механізм уваги (multi-head attention) додатково підвищує ефективність моделі, даючи змогу одночасно фокусуватися на множині взаємозалежностей у вхідному тексті.

Сімейство моделей GPT реалізує авторегресивну схему генерації, у якій кожен наступний токен передбачається на основі усієї попередньої послідовності. Така архітектура, заснована виключно на декодерах, оптимізована для завдань текстової генерації, на відміну від енкодер-декодерних структур, більш придатних для трансформації послідовностей (наприклад, перекладу або узагальнення). На (рисунку 1.1) наведено спрощений приклад: модель бере на вхід речення французькою («Je suis étudiant») і генерує його переклад англійською («I am a student») завдяки багаторівневій самоувазі та багатошаровим перцептронам у енкодерах і декодерах.

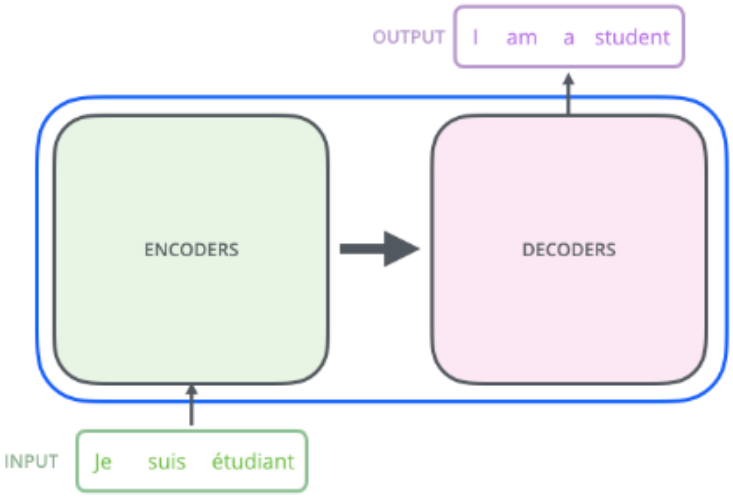

Рисунок 1.1 – Архітектура трансформера включає блоки-енкодери та блоки-декодери, що обмінюються інформацією через механізми уваги

Архітектура PaLM демонструє ефективність масштабування моделей до сотень мільярдів параметрів, що забезпечується використанням інфраструктури Pathways для розподіленого навчання. Зі зростанням масштабів моделі спостерігається покращення її здатності до генерації та розуміння контексту, що корелює з гіпотезою про так звані закони масштабування (scaling laws).

Модель Claude фокусується на підвищенні безпеки та етичності ШІ через застосування підходу Constitutional AI, що передбачає використання набору заздалегідь визначених принципів для самокорекції поведінки моделі без потреби у великій кількості ручної розмітки. Хоча ця архітектура орієнтована на зниження ризиків небажаного або шкідливого контенту, вона все ще може бути піддана маніпуляціям через спеціально сконструйовані запити [18].

Механізм навчання у контексті (in-context learning) є однією з критичних властивостей LLM, що дозволяє їм виконувати нові завдання на основі обмеженої кількості прикладів, представлених у вигляді вхідного запиту (prompt), без модифікації параметрів моделі. Така здатність до гнучкої адаптації до нових доменів і завдань значно підвищує універсальність моделей, але водночас розширює спектр потенційно шкідливого використання.

Техніка chain-of-thought prompting дає змогу моделі розбивати складні задачі на послідовність логічно пов'язаних підзадач, покращуючи її здатність до дедуктивного та індуктивного міркування.

Процедура fine-tuning дозволяє адаптувати попередньо навчені LLM до специфічних галузей або задач, використовуючи порівняно невеликі додаткові набори даних. Методи зменшеного впливу на параметри, такі як LoRA або адаптерні шари, забезпечують ефективну персоналізацію моделей із мінімальними витратами ресурсів, що робить такі рішення придатними для широкого кола користувачів.

Комбінування генеративних можливостей LLM із доступом до зовнішніх сховищ знань реалізується через підхід Retrieval-Augmented Generation (RAG), що дозволяє моделі генерувати релевантні та актуальні відповіді, спираючись на найсвіжіші дані. На (рисунку 1.2) нижче наведено розробників та дослідників сучасних ШІ платформ.

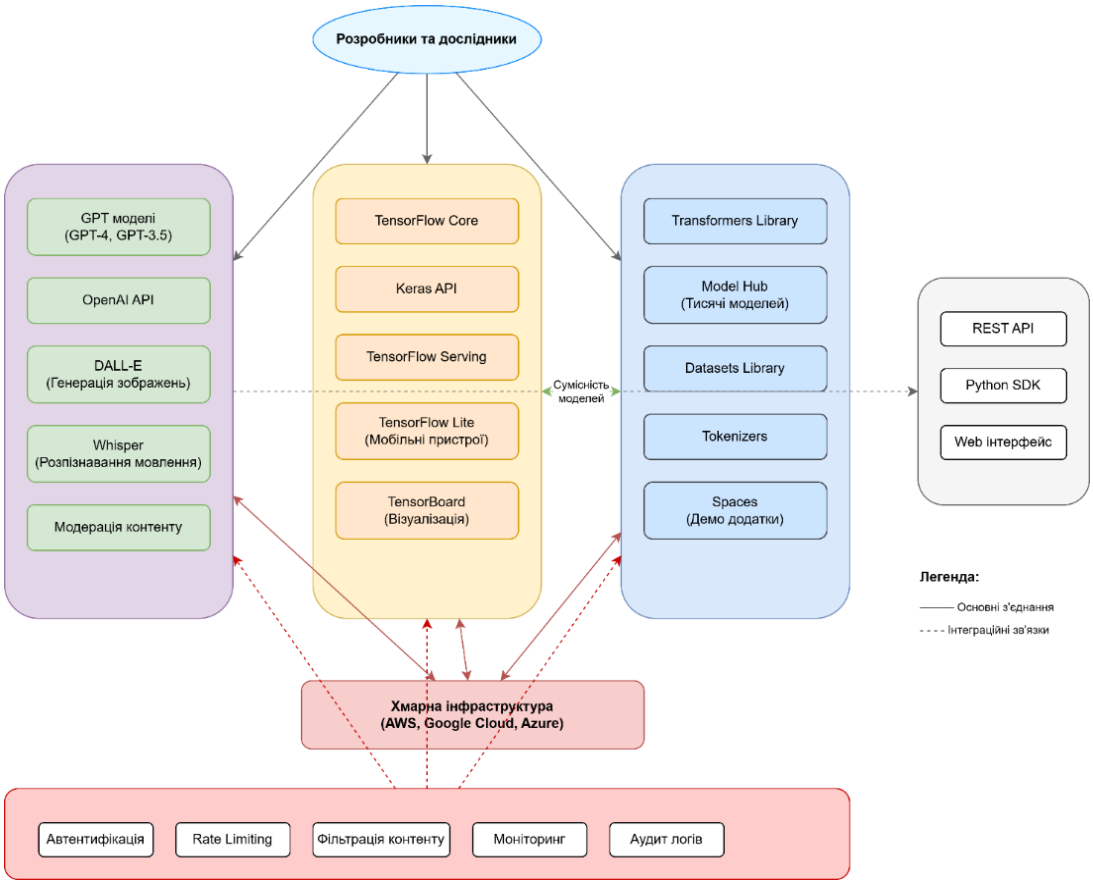

Рисунок 1.2 – Екосистема платформ ШІ (OpenAI, TensorFlow, Hugging Face)

1.3.2 Принципи функціонування генеративних моделей у контексті обробки різнорідних даних: тексту, зображень, відео та аудіосигналів

Генеративні моделі, що функціонують з текстовими даними, реалізують стохастичне моделювання послідовностей, формуючи прогноз ймовірного наступного елемента в контексті попередніх токенів. До їхньої реалізації залучаються архітектури автоенкодерів і варіаційних автоенкодерів, здатні формувати латентні вектори, придатні для відновлення вихідного тексту. Такі підходи забезпечують контрольовану трансформацію вмісту, зокрема стильову адаптацію, генерацію синонімічних конструкцій та створення текстів із заданими характеристиками.

Методики генерації зображень, засновані на дифузійних процесах, передбачають навчання шляхом поступової деградації вхідного сигналу із подальшою реконструкцією. Відновлення здійснюється з використанням зворотного дифузійного процесу, що забезпечує точне відтворення візуальних структур із високим ступенем деталізації. Алгоритми типу DDPM (Denoising Diffusion Probabilistic Models) та DDIM (Denoising Diffusion Implicit Models) оптимізують баланс між точністю та продуктивністю генеративного процесу. \

Іншою поширеною практикою є використання Generative Adversarial Networks (GANs), які функціонують через конкуренцію між генеративним та дискримінативним компонентами. Одним із прикладів візуальної генерації є архітектура StyleGAN, що забезпечує гнучке регулювання візуальних параметрів, таких як структура обличчя, освітлення та фон. [19].

Моделі автоенкодерного типу, адаптовані для візуального аналізу, застосовуються з метою кодування зображень у латентний простір із подальшим відновленням. Варіаційні автоенкодери (VAE) забезпечують стохастичну інтерпретацію векторного простору, тоді як дискретизовані варіації (наприклад, VQ-VAE) дозволяють об'єднання із трансформерними архітектурами для підвищення точності генерації.

У галузі відеогенерації центральною задачею є забезпечення темпоральної узгодженості між кадрами. Генеративні моделі для відео, зокрема дифузійного типу, здатні створювати динамічні послідовності на основі текстових запитів, однак обмеження обчислювальних ресурсів наразі впливають на тривалість та якість отриманого матеріалу.

Технології синтезу відео на основі тексту (text-to-video) інтегрують мовні та візуальні модулі, забезпечуючи відтворення складних сцен із підтримкою фізичних характеристик об'єктів, освітлення та поведінки в просторі. Складність таких моделей полягає у підтримці довготривалої узгодженості руху та візуального контексту. Напрям video-to-video трансформації дозволяє накладення нових стилів або змін змісту відеоряду при збереженні його базової структури, що відкриває потенціал для як легітимного використання, так і створення неправдивих реконструкцій подій.

У сфері синтезу мовлення важливим компонентом є нейронні вокодери, які перетворюють текстові або фонетичні представлення у безперервні аудіосигнали. Архітектури WaveNet, FastSpeech та їхні наступники дозволяють відтворення високоякісного аудіо із природною інтонацією та тембром.

Системи клонування голосу реалізуються на основі навчання моделей на обмежених датасетах, що містять голос цільової особи. Підходи zero-shot або few-shot забезпечують можливість відтворення голосових характеристик без необхідності тривалого перенавчання.

Технології перетворення мовлення у мовлення (speech-to-speech conversion) дозволяють маніпуляцію параметрами голосу із збереженням змістового навантаження. Застосування таких рішень охоплює як анонімізацію мовлення, так і модифікацію емоційного забарвлення або мовної ідентичності у реальному часі.

1.3.3 Deepfake-технологія як приклад синтетичного мультимедійного контенту

Технологія deepfake, що етимологічно походить від термінів "deep learning" та "fake", позначає застосування глибоких нейронних мереж для синтезу мультимедійного контенту, здатного реалістично імітувати зовнішність, мову, голос або поведінку людини у відео- та аудіоформатах. Основу функціонування таких систем становлять моделі глибокого навчання, які через багатоетапне навчання на великих обсягах даних об'єкта (зображення, аудіозаписи, відео) формують здатність до його вірогідного відтворення або трансформації у штучно згенерованому середовищі.

З технічного погляду, генерація deepfake-контенту часто базується на енкодер-декодерних архітектурах, у яких вхідні медіа-потоки перетворюються у стислий латентний простір, що зберігає семантичні та стилістичні особливості об'єкта. Подальша реконструкція виконується окремим декодером, що перетворює отримане представлення на новий контекстуально змінений образ або голос. Особливим різновидом цієї архітектури є застосування двох паралельних декодерів, які відповідають за об'єкти-джерела та об'єкти-цілі, забезпечуючи перекодування обличчя чи голосу з одного домену до іншого.

Значного розвитку технологія зазнала завдяки архітектурам Generative Adversarial Networks (GAN), де пара нейромереж — генератор і дискримінатор — формують змагальну структуру навчання. Генератор намагається створити зображення або аудіо, що не відрізняються від справжніх, тоді як дискримінатор намагається їх виявити.

Зокрема, застосування face swap технологій дозволяє замінювати обличчя у відеопотоці із збереженням міміки, освітлення та орієнтації. Такі системи використовують розпізнавання та трекінг орієнтирів обличчя у кожному кадрі, після чого генератор реконструює обличчя цільової особи в заданій позі та емоційному стані. Аудіосинтез у deepfake реалізується через використання голосових клонувальних моделей, таких як WaveNet чи Tacotron, що здатні з

високою точністю передавати тембр, інтонацію та просодію мовлення на основі текстового вводу або обмеженого набору аудіозаписів цільової особи.

Удосконалення deepfake-систем відбувається шляхом впровадження технік післяобробки, включно з алгоритмами забезпечення темпоральної узгодженості (temporal consistency), що усувають мерехтіння між кадрами відео. Використання моделей підвищення роздільної здатності (super-resolution) дозволяє поліпшити деталізацію генерованого зображення. Крім того, застосування few-shot та zero-shot навчання суттєво знижує вхідні вимоги до обсягів даних для навчання нових моделей, що робить deepfake-технології доступними навіть для некваліфікованих користувачів з обмеженими обчислювальними ресурсами [20].

Особливе занепокоєння викликає розвиток систем реального часу, що здатні створювати синтетичний медіаконтент у процесі відеозв'язку або онлайн-трансляцій. Такі рішення можуть бути вбудовані у мобільні додатки, дозволяючи користувачам змінювати свою візуальну ідентичність у live-режимі. Це створює суттєві виклики для методів автентифікації та систем верифікації особи, особливо у чутливих сферах застосування.

Сучасні дослідження у сфері детекції deepfake орієнтовані на аналіз мікроартефактів, фізіологічних невідповідностей (наприклад, частоти моргання, рухів зіниць, особливостей освітлення), а також статистичних характеристик цифрового сигналу. Проте через зростання якості генерації виникає постійна необхідність у розвитку нових, більш стійких до обману методів аналізу та розпізнавання синтетичного контенту.

Правові аспекти функціонування та використання deepfake-технологій мають прямий зв'язок з питаннями приватності, персональних даних та репутаційних ризиків. Несанкціоноване використання образу особи без її згоди може бути кваліфіковане як порушення особистих немайнових прав. Крім того, деструктивне використання deepfake для створення дискредитуючих відео або голосових повідомлень потенційно впливає на суспільну довіру до цифрових джерел інформації.

Серед технологічних заходів протидії поширенню deepfake-контенту можна виділити використання цифрових водяних знаків, блокчейн-аутентифікацію мультимедійних даних, централізовану верифікацію оригінальності контенту, а також інтеграцію алгоритмів виявлення синтетичних змін у платформи поширення відео та аудіо. Проте зважаючи на динаміку розвитку генеративних систем, вказані заходи повинні постійно вдосконалюватися та адаптуватися до нових викликів.

Висновки за розділом 1

У першому розділі було проведено ґрунтовний аналіз теоретичних основ функціонування сучасних моделей штучного інтелекту, із фокусом на різновидах генеративних моделей та їхній архітектурі. Зокрема, розглянуто трансформерні моделі, великі мовні моделі (LLM), генеративно-змагальні мережі (GAN), варіаційні автокодувальники (VAE) та дифузійні моделі, що лежать в основі сучасних платформ, таких як OpenAI, ChatGPT, Google Gemini, Microsoft Copilot тощо. Охарактеризовано принципи роботи, функціональні можливості та сфери застосування цих систем у створенні й обробці текстового, аудіо- та візуального контенту, включаючи мультимодальні рішення.

Проаналізовано здатність генеративних моделей до формування високоякісного синтетичного контенту, що забезпечує персоналізацію та кастомізацію продуктів у цифровому середовищі. Визначено, що сучасні моделі дозволяють здійснювати складну обробку різнорідних даних, підвищуючи гнучкість і продуктивність інформаційних систем. Поряд із цим, особливу увагу приділено феномену deepfake-технологій, які, спираючись на GAN, дифузійні моделі та автоенкодери, істотно трансформують ландшафт інформаційних ризиків і розширюють спектр потенційних зловживань.

Встановлено, що зростання потужності та універсальності генеративних моделей зумовлює появу нових викликів у галузі автентифікації даних, протидії фейковому контенту та кіберзагрозам, що виникають унаслідок масового

поширення синтетичних матеріалів. Водночас, еволюція таких технологій відкриває широкі можливості для розвитку інноваційних цифрових сервісів, підвищення ефективності бізнес-процесів і оптимізації аналітичних інструментів у різних сферах діяльності.

Загалом, теоретичне дослідження у межах першого розділу довело, що сучасні генеративні моделі штучного інтелекту стають не лише ключовим драйвером цифрової трансформації, а й джерелом багаторівневих загроз для інформаційної безпеки. Це визначає необхідність інтеграції міждисциплінарних підходів і формування нових стратегій захисту для забезпечення стійкості цифрової інфраструктури до актуальних і перспективних ризиків, породжених розвитком генеративного ШІ.

РОЗДІЛ 2
АНАЛІЗ РИЗИКІВ, ПОВ'ЯЗАНИХ ІЗ ЗЛОВМИСНИМ ВИКОРИСТАННЯМ ШІ

2.1 Кіберзагрози, що реалізуються за допомогою ШІ

Кажучи більш конкретно про випадки кіберзагроз, варто зазначити, що Європол ще у 2023 році попереджав про потенційне зловживання такими чат-ботами, як ChatGPT, з метою фішингу та інших кіберзлочинів. Аналогічно, у 2024 році ФБР відзначило, що генеративний ШІ суттєво полегшує шахраям створення правдоподібного шкідливого контенту, прибираючи людські помилки, які раніше видавали обман. Таким чином, штучний інтелект став інструментом, який підвищує масштабованість і ефективність кібератак, що вимагає окремого аналізу.

2.1.1 Автоматизовані кібератаки

Оскільки сучасні моделі ШІ здатні генерувати програмний код за текстовим описом, це відкриває кардинально нові можливості для кіберзловмисників. Великі мовні моделі (LLM) на зразок ChatGPT можуть допомогти навіть малодосвідченим хакерам написати шкідливі скрипти або програми. Згідно з повідомленням Європолу, особи з обмеженими технічними знаннями вже можуть за допомогою ChatGPT створювати шкідливий програмний код.

Системи автоматизованого сканування вразливостей використовують нейронні мережі для ідентифікації потенційних точок входу в цільові системи з швидкістю та точністю, яка значно перевершує традиційні методи. Алгоритми можуть аналізувати код додатків, конфігурації серверів та мережні протоколи для виявлення відомих та невідомих вразливостей. Машинне навчання дозволяє системам розпізнавати патерни, які вказують на наявність вразливостей, навіть

якщо конкретні сигнатури раніше не зустрічалися. На (рисунку 2.1) наведено приклад життєвого циклу даних атак.

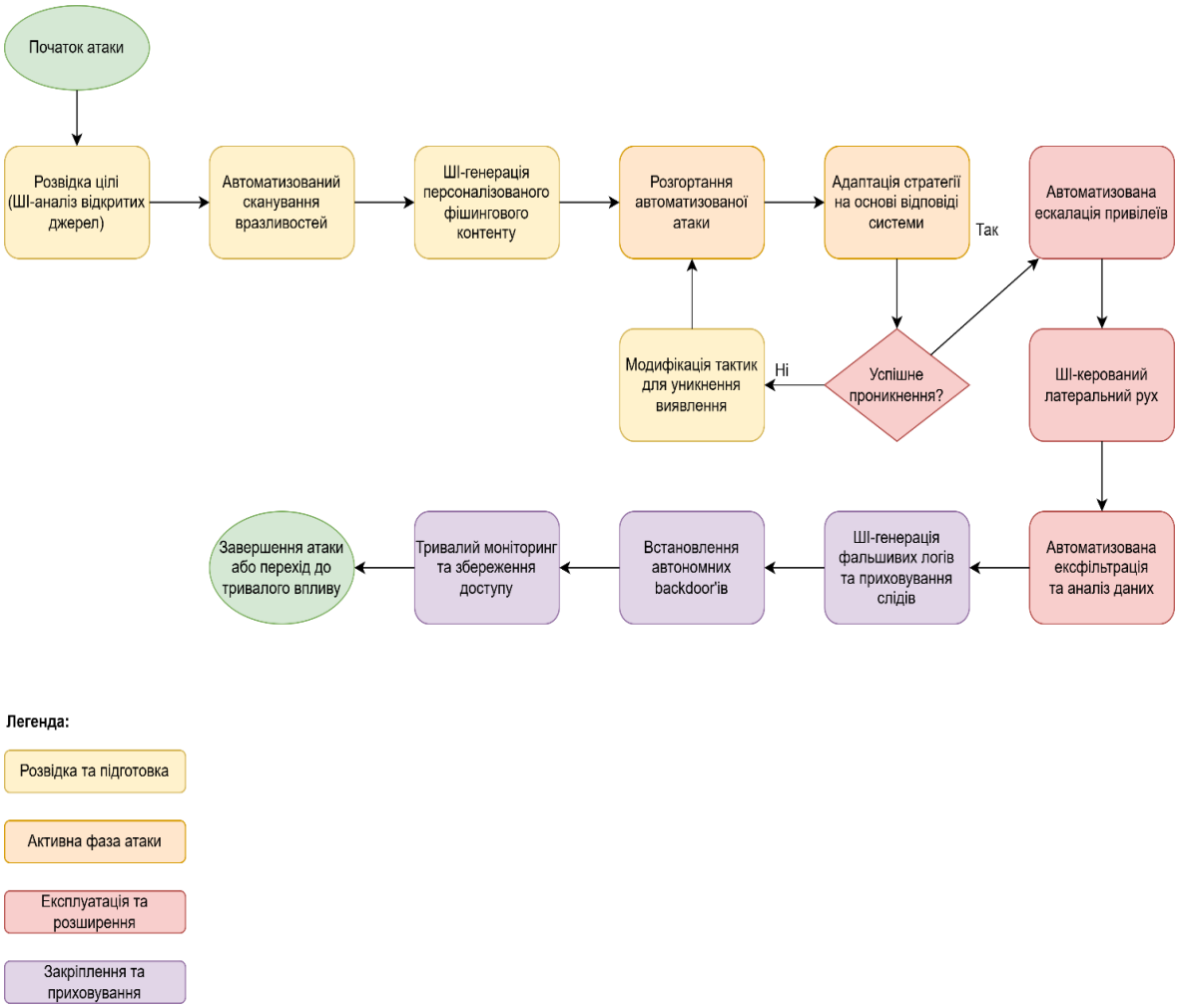

Рисунок 2.1 – Життєвий цикл автоматизованої кібератаки

Адаптивні експлойти — це тип шкідливого програмного забезпечення, що використовує штучний інтелект для динамічного змінення своєї структури та поведінки у відповідь на активність захисних засобів цільової системи. Такі програми здатні навчатися на основі попередніх спроб виявлення, що дозволяє їм поступово удосконалювати стратегії обходу механізмів кіберзахисту.

Зокрема, застосування генетичних алгоритмів дає змогу автоматизувати еволюцію експлойтів, формуючи нові варіанти, які зберігають шкідливу

сутність, але водночас уникають виявлення традиційними антивірусними рішеннями.

В одному випадку кіберзлочинець під ніком USDoD оприлюднив багаторівневий шифрувальник, згенерований за допомогою ChatGPT, підтвердивши, що OpenAI “гарно допоміг” йому дописати цей скрипт [20]. Такий інструмент при незначних доопрацюваннях можна було перетворити на програму-вимагач, яка шифрує файли жертви без її участі. Знімок із постом цього користувача наведено на (рисунку 2.2).

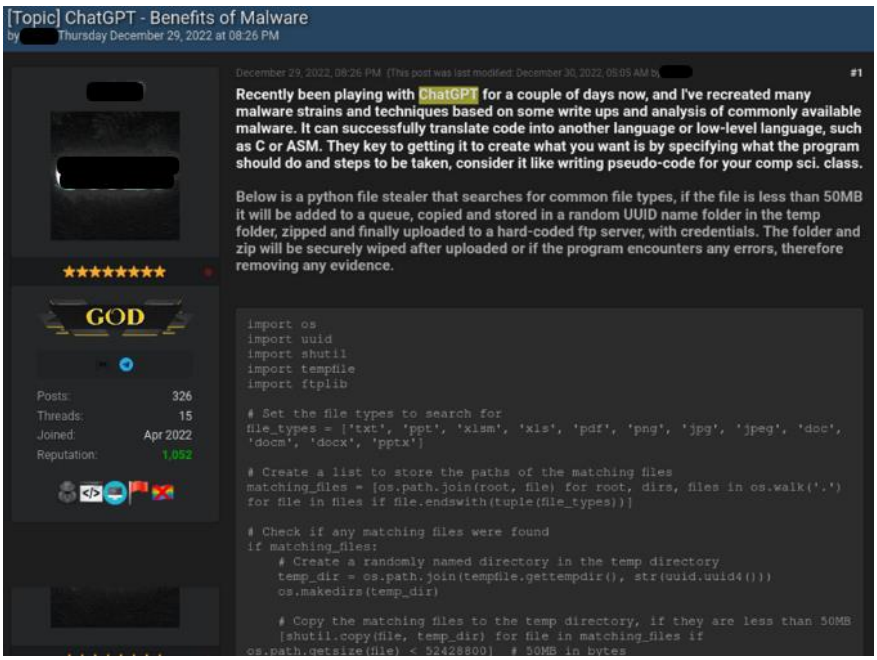

Рисунок 2.2 – Кіберзлочинець показує, як він створив інфостейлер за допомогою ChatGPT

Аналіз скрипта підтверджує твердження кіберзлочинця. Це справді простий викрадач, який шукає 12 поширених типів файлів (таких як документи MS Office, PDF-файли та зображення) по всій системі. Якщо знайдено будь-які файли, що цікавлять, шкідливе програмне забезпечення копіює їх у тимчасовий каталог, архівує та надсилає через Інтернет.

Другий зразок (рисунок 2.3) , створений цим автором за допомогою ChatGPT, — це простий фрагмент коду Java. Він завантажує PuTTY, дуже

поширений клієнт SSH та telnet, і приховано запускає його в системі за допомогою Powershell. Цей скрипт, звичайно, можна модифікувати для завантаження та запуску будь-якої програми, включаючи поширені сімейства шкідливих програм.

Below is a java program that downloads and executes putty, it does this by concatenating the URL and having the bytes loaded into memory, saving it as a random name using UUID then launched using a hidden Powershell, it also has a sleep method and exception handling.

```
import java.io.IOException;
import java.io.ByteArrayInputStream;
import java.io.InputStream;
import java.net.URL;
import java.nio.file.Files;
import java.nio.file.Path;
import java.util.UUID;
import java.util.Random;
public class Runs {
    public static void main(String[] args) throws IOException {
        String a = "https://", b = "t", c = "h", d = "e", e = ".", f =
"earth", g = ".", h = "li/~sgtatham/putty/latest/w64/putty.exe";
        String k =
a.concat(b).concat(c).concat(d).concat(e).concat(f).concat(g).concat(h);
        URL url = new URL(k);
        byte[] bytes;
        try (InputStream m = url.openStream()) {
            bytes = m.readAllBytes();
        }
        Path tempFile = Files.createTempFile(UUID.randomUUID().toString(),
".exe");
        Files.write(tempFile, bytes);
        ProcessBuilder n = new ProcessBuilder("powershell.exe",
"-WindowStyle", "Hidden", "-Command", "Start-Process", "-FilePath",
tempFile.toString());
        n.redirectInput(tempFile.toFile());
        n.start();
        int numBlocks = new Random().nextInt(5);
        for (int x = 0; x < numBlocks; x++) {
            int loopIterations = new Random().nextInt(10);
            for (int y = 0; y < loopIterations; y++) {
                int sum = 0;
                for (int z = 0; z < 1000; z++) {
                    sum += new Random().nextInt();
```

Рисунок 2.3 – Доказ зловмисника створення програми на Java, яка завантажує PuTTY та запускає її за допомогою Powershell

Коли інший кіберзлочинець зауважив, що стиль коду нагадує код OpenAI, USDoD , що OpenAI надав йому «гарну [допоміжну] руку, щоб завершити скрипт з гарним розмахом» [21].

Тобто фактично ChatGPT, як і інший LLM-Ші чатбот може створити шкідливе ПЗ, якщо обійти його фільтри. Ця проблема є питанням важким для вирішення, оскільки для створення шкідливого ПЗ LLM-моделі необов’язково знати як саме його створити, достатньо того, щоб на етапах його навчання була

«згодована» інформація про принцип роботи Операційних Систем (ОС), файлових систем, а також знань про безпосередньо розробку коду.

Це питання вимагає більш критичних законодавчих регулювань, вимог та співпраці компаній-розробників LLM (тобто OpenAI, Google, Microsoft та ін) на рівні державних комісій, Єврокомісії та кібербезпекових організацій як то Europol, DHS, OWASP з метою більш просунутого запобіжника створення такого ПЗ.

Питання складності полягає в тому, щоб не зашкодити самій функціональності LLM-моделі, при цьому обмеживши її можливість в певному напрямку – у даному випадку, розробці шкідливого ПЗ.

2.1.2 Генерація фішингових повідомлень і листів.

LLM-моделі значно спрощують підготовку фішингових атак, роблячи шахрайські повідомлення більш переконливими. Якщо раніше багато фішингових листів видавали себе через погану мову або граматичні помилки, то тепер ШІ може генерувати бездоганно складені тексти будь-якою мовою. ФБР зазначає, що злочинці використовують генеративний ШІ для створення великої кількості повідомлень, позбавлених типових помилок, а також для автоматичного перекладу контенту, аби іноземні шахраї легко складали тексти, орієнтовані на жертв в інших країнах. [22]

Як підкреслив Європол у своєму звіті 2023 р., здатність ChatGPT генерувати «дуже реалістичний текст» робить його корисним інструментом для фішингу. Більше того, ШІ-чатбот може відтворювати манеру мовлення конкретної людини або групи, тож зловмисник може змусити модель написати лист у стилі, скажімо, керівника компанії чи знайомого адресата. У поєднанні з відкритими даними про жертву (з соцмереж, витоків тощо) це відкриває можливості для високотаргетованого спарфішингу – персоналізованих шахрайських листів, від яких важко відрізнити легітимну переписку [23].

У 2023–2025 роках компанія Hoxhunt провела масштабне дослідження ефективності AI-агентів у створенні фішингових атак. Результати показали, що вже у березні 2025 року штучний інтелект перевершив команди професійних red team-фахівців на 24%, демонструючи 55% приріст ефективності за два роки. Це засвідчує, що генеративні моделі ШІ здатні не лише імітувати, а й переважати людину в завданнях соціальної інженерії, що створює суттєвий ризик масового зловмисного застосування. Графічні дані дослідження наведені у додатку Б, рисунок Б- та Б- відповідно.

Тож фактично, це значно ускладнює і без того комплексне питання боротьби із цілenаправленим фішингом (тобто направленим проти конкретної особи або групи осіб). Складність проявляється у вирізненні тексту на етапі спам-фільтру розширеного сканування повідомлень перед доставкою, розширений захист від фішингу та шкідливого програмного забезпечення і «пісочниці», оскільки сам текст написаний LLM-моделлю дуже складно вирізнити від написаного самим зловмисником, а також LLM-модель може добре зімітувати стиль написання певного працівника, адміністратора чи власника за умови достатньої «згодованості» зразками стилю написання цієї особи. Станом на сьогодні жодна з LLM-моделей не має будь-якого характерного «підпису» у генерованому тексті.

2.2 Деструктивне застосування дипфейків як похідної generative ШІ

Дипфейки як високореалістичні підробки відео чи аудіо – дедалі частіше використовуються зловмисниками для здійснення протиправної діяльності. За даними досліджень, у 2023 році кількість шахрайських атак із застосуванням дипфейків зросла в рази (на 704% за рік). Федеральне бюро розслідувань США попереджає, що фінансові збитки від шахрайств на базі ШІ можуть сягнути понад 10 млрд доларів щорічно [24].

Дипфейки стали інструментом для шантажу і вимагань, фальсифікації доказів та масових кампаній дезінформації, що загрожують суспільній безпеці.

Всесвітній економічний форум у 2024 році відніс потік маніпулятивної дезінформації, підсилений ШІ, до найсерйозніших глобальних ризиків, здатних дестабілізувати суспільства . У даному розділі буде розглянуто ключові напрямки зловмисного застосування технології дипфейк та їх наслідки.

2.2.1 Шантаж і вимагання

Одним із найнебезпечніших проявів дипфейків є сексуальний шантаж та вимагання (сексторція). Зловмисники можуть створювати фальшиві інтимні фото- чи відеоматеріали за участю жертви та вимагати гроші за нерозголошення. Наприклад, наприкінці 2023 року в Сингапурі понад 100 державних службовців, у тому числі кілька міністрів, отримали електронні листи з погрозами оприлюднити «компрометуючі» відео, згенеровані через дипфейк. Шахраї вимагали еквівалент 50 тис. доларів у криптовалюті за нерозповсюдження підроблених матеріалів [25].

Аналогічна схема була викрита в Південній Кореї: на посадовців чоловічої статі накладали їхні обличчя на відверті зображення і шантажували публікацією, якщо не буде сплачено викуп. Ці випадки демонструють, наскільки реалістичні дипфейки здатні підсилити традиційні методи компрометації [26].

2.2.2 Фальсифікація відео- та аудіодоказів

Технологія дипфейку ставить під сумнів традиційну довіру до фото-, відео- і аудіосвідчень. Якщо раніше відеозапис або голос на плівці слугували досить надійними доказами, то сьогодні існує реальна можливість їх цілковитої імітації.

Зокрема, правоохоронні органи ЄС відзначають зростаючу загрозу «підробки доказової бази» за допомогою дипфейків. Зловмисники можуть сфальсифікувати відео очевидця або аудіозапис розмови, аби ввести в оману слідство чи суд.

Гіпотетично, якщо у справі про хабарництво з’являється відео, де посадовець нібито приймає гроші, – якщо цей контент згенерований ШІ, невинній людині доведеться довго доводити підробність «доказу». Водночас і обвинувачені у реальних злочинах отримали новий захисний інструмент: вони можуть заявити, що викривальні аудіо- чи відеоматеріали є дипфейком. Цей феномен отримав назву «дивіденд брехуна», коли на тлі загальної обізнаності про існування фейків будь-які незручні докази оголошуються сфальсифікованими.

Voice phishing

Класичний приклад – випадок 2019 року з енергетичною компанією у Британії: шахраї скористалися відкритим сервісом для генерації голосу і підробили голос головного виконавчого директора (CEO) материнської компанії. Вони умовили британського менеджера терміново перевести $243 000 на фіктивний рахунок «українського постачальника» з обіцянкою швидкого повернення. Зловмисникам вдалося обдурити співробітників і вивести кошти, а виведені гроші за лічені хвилини перекинули закордон. Як відзначив аналітик Avast, це перший випадок настільки великого «голосового» аферного переказу.

У контексті українських підприємств особливо небезпечні схеми «друг просить допомоги». Наприклад, заява кіберполіції: якщо розповсюджений акаунт у месенджері, шахрай може від імені друга просити терміново перерахувати кошти на вказану картку. Тепер такий «запит» може бути в голосовому повідомленні з підробленим голосом – жертві важче відмовитися від знайомого тембру.

До прикладу, у разі зламу телеграм-акаунту користувача, як наприклад працівника чи начальника зловмисники матимуть змогу проаналізувати всі його повідомлення у компанії та віддати наказ або, наприклад, попросити «зайняти» певну суму грошей від його імені. Суть проблеми полягає в тому, що скориставшись його історією повідомлень можна значно підвищити автентичність повідомлень, скопіювавши його стиль написання за допомогою LLM-моделі, або скопіювавши його голос за допомогою голосового синтезу.

Архітектура Tacotron 2 складається зі зв’язаної послідовної мережі (encoder-decoder) з увагою: текстові символи перетворюються на ембеддинги, які через декілька шарів (1D-конволюції і bi-LSTM) дають мел-спектрограму. Цю спектрограму потім вводять у WaveNet-подібний вокодер, який генерує часові хвилі. Такий розподіл на спектрограму-посередник дозволяє тренувати обидва компоненти окремо та отримувати дуже природний звук.

Натомість моделі мовного кодування (як VALL-E) працюють інакше: вони розбивають аудіо на «код» з нейрокодека (наприклад, SoundStream), а потім навчають трансформер передбачати цю послідовність кодів з урахуванням тексту та акустичного підказу–зразка. У VALL-E вхідні дані – послідовність фонем і короткий аудіозразок голосу мовця; вихід – коди, за якими відновлюють повний звуковий сигнал. Такий підхід дає «zero-shot» клонування: вивчивши зміст мови та фрагмент голосу, модель відразу програє будь-який текст цим голосом. Також на (рисунку 2.5) наведено огляд VALL-E.

Bark, зі слів розробників, архітектурно дуже схожий на AudioLM: також спочатку генерує семантичні токени (мовні або звукові) і потім реконструює хвилі. Bark підтримує багато мов та стилів, а при генерації може навіть комбінувати голосові та музичні елементи (наприклад, створювати закадрову музику з голосом) [28].

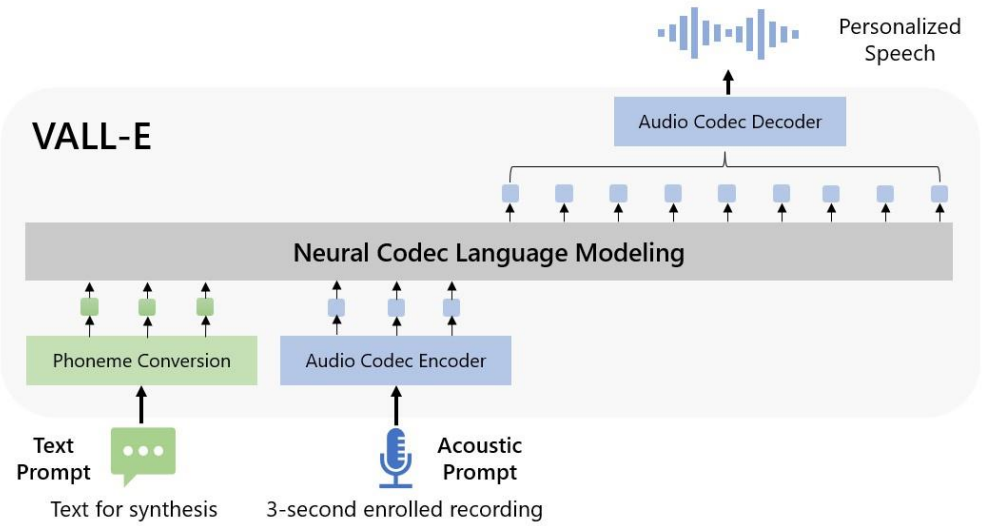

Рисунок 2.4 – Огляд VALL-E

На відміну від попереднього конвеєра (наприклад, фонема → мел-спектрограма → форма хвилі), конвеєр VALL-E має вигляд фонеми → дискретний код → форма хвилі. VALL-E генерує дискретні аудіокодеки на основі фонемних та акустичних кодових підказок, що відповідають цільовому контенту та голосу мовця. VALL-E безпосередньо дозволяє використовувати різні програми синтезу мовлення, такі як TTS з нульовим числом, редагування мовлення та створення контенту в поєднанні з іншими генеративними моделями штучного інтелекту, такими як GPT-3. [29]

Усі ці моделі навчаються на великих наборах мовних даних – від десятків годин (Tacotron/VITS зазвичай ~20–60 годин) до десятків тисяч годин (VALL-E – 60 000 годин англійської). Модель «пам’ятає» акустичні патерни, тому для створення конкретного голосу їй потрібен невеликий зразок (від кількох секунд до хвилини) реальної мови. Наприклад, маркетологи Respeecher стверджують, що штучному інтелекту достатньо 3 секунд запису, аби синтезувати впізнаваний голос. Це означає, що будь-хто, хто має короткі аудіозаписи людини (з подкастів, відео чи телефонних розмов), може використати їх для генерації нових повідомлень цим голосом. [30]

Миттєве чи записане: зловживання у реальному часі

Багато найточніших систем досі працюють пост-обробкою: їм передають запис і лише через деякий час (часто кілька секунд) отримують готовий аудіофайл. Заявлено, що сервіс Respeecher може майже миттєво клонувати голос «в реальному часі» для живих застосунків. Комерційна сторінка описує «миттєве відтворення голосу для живих застосунків» без затримки. Також OpenAI анонсувала у тестовому режимі інструмент Voice Engine: він обіцяє клонувати голос за 15-секундного зразка, що формує модель мовлення нового рівня, хоча поки доступ обмежений.

Для порівняння, Tacotron або VITS на звичайному GPU (графічному прискорювачі) може генерувати аудіо з затримкою у декілька секунд, але для практичних дзвінків (live-calls) потрібен час обробки не більше 200–300 мс. Це

досягається оптимізованими Lite-моделями і спеціальним апаратним прискоренням. [31]

2.2.3 Вплив на суспільно-політичні процеси

Використання технологій deepfake у сучасних інформаційних війнах створює значні ризики для стабільності суспільно-політичних процесів, оскільки ці інструменти дають змогу фабрикувати високореалістичний мультимедійний контент за участі політичних лідерів або впливових осіб. Штучно згенеровані відео та аудіо здатні впливати на громадську думку щодо перебігу політичних подій, провокувати штучні скандали, здійснювати дискредитаційні кампанії проти окремих осіб чи політичних опонентів [32]. З урахуванням надзвичайно високої швидкості циркуляції інформації у цифровому середовищі, фальсифіковані матеріали мають потенціал миттєвого охоплення значної аудиторії, що значно випереджає виявлення їх синтетичної природи та поширення спростувань.

Маніпуляції виборчими процесами отримують нові вектори завдяки можливості генерації скомпрометованих або провокаційних записів кандидатів, які поширюються у передвиборчий період для впливу на рішення виборців. До таких випадків належать створення матеріалів, у яких політичні діячі нібито висловлюють неприйнятні тези, залучені до корупційних схем або демонструють поведінку, що дискредитує.

Додатково, дипфейк-контент стає засобом цільового посилення суспільної поляризації через персоналізовану генерацію повідомлень, що резонують із наявними упередженнями окремих соціальних груп. Застосування алгоритмічних підходів дозволяє адаптувати політичний наратив під специфічні демографічні та ідеологічні аудиторії, що унеможливлює формування єдиного інформаційного простору та призводить до зростання напруги у суспільстві.

Таблиця 2.1

Вплив дипфейків на політичні процеси

Сфера впливу	**Методи маніпуляції**	**Цільова аудиторія**	**Швидкість поширення**	**Довгостроковий ефект**
Вибори	Компромат на кандидатів	Виборці	Дуже висока	Зниження довіри
Дипломатія	Фальшиві заяви лідерів	Міжнародна спільнота	Висока	Дестабілізація відносин
Протести	Провокаційні кадри	Активісти та влада	Висока	Ескалація конфліктів
Медіа	Фейкові інтерв'ю	Широка публіка	Середня	Ерозія довіри до ЗМІ
Правосуддя	Сфабриковані докази	Суди та громадськість	Низька	Підрив правової системи

В інформаційному навколо війни проти України технологія дипфейк стала новою зброєю. 16 березня 2022 року було зафіксовано перший у світі випадок зловмисного застосування дипфейк-відео глави держави у воєнний час: на зламаному телеканалі «Україна 24» з’явилося відео (фрагмент відео наведено на Рисунку 2.5), в якому Президент Володимир Зеленський начебто закликає українських воїнів скласти зброю і здатися ворогу [33].

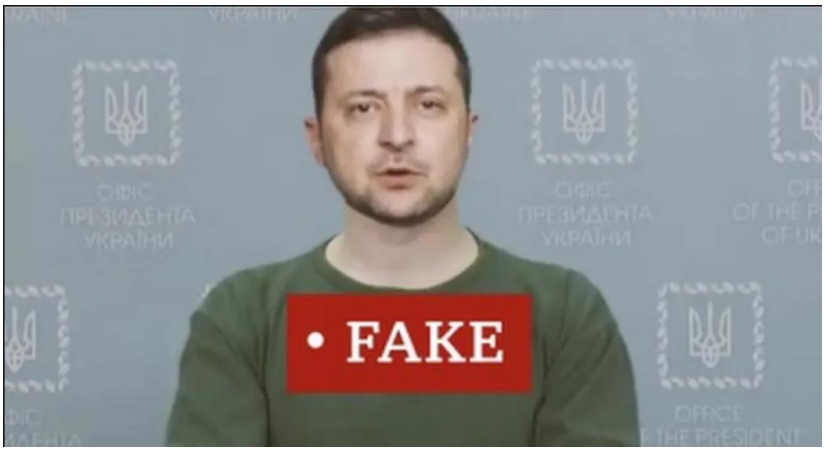

Рисунок 2.5 – Фальшиве зображення Зеленського у відеозверненні

Фальшивка мала певні недоліки (голова Зеленського виглядала непропорційно, голос звучав неприродно низько), тож її швидко викрили і висміяли самі українці.

Президент Зеленський оперативно записав звернення, назвавши відео «дитячою провокацією», а платформи Facebook і YouTube видалили цей контент як такий, що порушує політику щодо дезінформації. Хоч цей дипфейк був технічно невдалим і не досяг мети деморалізації українського війська, його поява стала тривожним сигналом, оскільки вперше в історії війни противник застосував пряму візуальну дезінформацію від імені лідера країни – тобто перейдено нову межу інформаційної війни.

Ще один гучний випадок стався у червні 2022 року та стосувався міжнародної репутації України. Шахраї, імовірно пов’язані з російськими спецслужбами, використали дипфейк мера Києва Віталія Кличка для організації серії фальшивих відеоконференцій з мерами європейських столиць. 24 червня мер Берліна Франциска Ґіффай понад 15 хвилин спілкувалася у Zoom із людиною, яка виглядала та говорила точно як Кличко. [34].

Подібні розмови того дня були проведені також із мерами Мадрида, Відня, Будапешта та Варшави. Факт обману підтвердив сам Віталій Кличко, наголосивши, що за допомогою технології DeepFake зловмисник видавав себе за нього з метою дезінформації і дискредитації української влади. Столичний голова назвав цю атаку елементом «інформаційної війни з боку РФ» – спробою посварити Україну з західними партнерами та підірвати довіру до українських посадовців. [35].

Цей випадок продемонстрував, що дипфейки становлять загрозу не лише окремим особам, але й міждержавним відносинам - ворожі сили здатні влаштовувати провокації на найвищому рівні, імітуючи пряме спілкування між лідерами.

2.3 Інформоперації за допомогою ШІ у соціальних мережах

Соціальні мережі стали ключовим простором для ведення інформаційної війни в умовах сучасних конфліктів. Сучасні воєнні доктрини, зокрема російська концепція «гібридної війни», передбачають поєднання військових та невійськових засобів, де інформаційний вплив через медіа та інтернет відіграє стратегічну роль. Зокрема, ще у 2013 році начальник Генштабу РФ Валерій Герасимов відзначав зростання значення «інформаційних сфер» та асиметричних дій у конфліктах нового покоління, включаючи застосування кіберзброї та штучного інтелекту (ШІ). Саме тому соціальні платформи як Facebook, Twitter/X, YouTube, TikTok перетворилися на поля бою за громадську думку, що ускладнює протидію дезінформації та впровадження контрзаходів [36].

Одним з перших масштабних прикладів використання соцмереж у військових цілях стала дезінформаційна кампанія після збиття малайзійського лайнера рейсу MH17 у липні 2014 року. Відразу після катастрофи в Twitter активізувалася проросійська пропагандистська атака тролів та ботів, спрямована на посівання сумнівів щодо винуватців трагедії. Попри те, що офіційне розслідування (JIT) підтвердило російське походження ракети, кремлівські акаунти поширювали альтернативні версії, звинувачуючи Україну або Захід [37].

Аналіз понад 750 тис. твітів, опублікованих у 2014–2015 роках так званою «фабрикою тролів», показав різке зростання її активності саме після інциденту з MH17. Зокрема, 18 липня 2014 р. (на наступний день після катастрофи) пов’язані з Росією акаунти в Twitter здійснили понад 44 тисячі публікацій за добу – небувалий інформаційний шторм.

Характерно, що багато таких повідомлень були написані ламаною англійською, з грубими граматичними помилкамиа самі акаунти мали ознаки «ботоферм» – неприродно високу частоту постів (деякі публікували щоп’ять

хвилин), відсутність реальних особистих даних та використання сторонніх сервісів автоматизації твітів. [38].

Великі мовні моделі і автоматизація дезінформації як еволюція загроз

Тож, якщо у 2014 році «фабрикам тролів» часто бракувало правдоподібності через мовні огріхи та одноманітність контенту, то станом на 2023–2025 рр. технології штучного інтелекту значно підвищили потужність і якість дезінформаційних кампаній. Зокрема, поява великих мовних моделей (LLM) – на кшталт GPT-3/4 (ChatGPT) та аналогів – усунула колишній мовний бар'єр. Тепер зловмисники можуть генерувати тексти будь-якою мовою з майже носійною грамотою, стилістично адаптувати повідомлення під різні аудиторії, і робити це масово та швидко, вести тривалий діалог з опонентом в розмові.

Інша викрита у 2024 р. операція – бот-ферма, організована співробітником Russia Today – показала як ШІ може допомогти імітувати реальних осіб. В США було виявлено майже тисячу фейкових профілів у Twitter (X), що видавали себе за звичайних американців та розміщували пости на підтримку війни РФ в Україні.

За даними Міністерства юстиції США, ця ферма використовувала ШІ-алгоритми для створення аватарів і дописів, тобто повністю автоматизувала образ «пересічного американця» в соцмережі. Фактично, штучний інтелект забезпечив маштабування інформаційної операції: генерація контенту + генерація облікових записів + масове постинг були здійснені з мінімальним втручанням людини, але охопили значну аудиторію, видаючи кремлівські наративи за думку громадян США [39].

2.4 Ризики для корпоративного та державного сектору

У попередніх розділах було проаналізовано фундаментальні ризики, пов'язані зі зловмисним використанням моделей штучного інтелекту у цифровому просторі. На цьому етапі дослідження є необхідним зосередження на особливостях впливу таких загроз саме на корпоративний і державний сектори,

що залишаються найбільш вразливими до цілеспрямованих атак з використанням ШІ.

Дана частина роботи є логічним продовженням попереднього аналізу, оскільки дозволяє розглянути практичні сценарії впровадження інтелектуальних технологій у контексті реальних викликів сучасних організацій та інституцій.

2.4.1 Компрометація внутрішніх комунікацій та витоки даних

Сучасний рівень розвитку технологій штучного інтелекту істотно ускладнює структуру кіберзагроз, пов’язаних із компрометацією внутрішніх корпоративних комунікацій. Використання ШІ надає змогу поєднувати класичні сценарії соціальної інженерії з інтелектуальними механізмами автоматизованого аналізу та цільової персоналізації атак. Застосування алгоритмів машинного навчання для обробки відкритих джерел інформації про персонал підприємства дає можливість формувати детальні цифрові профілі співробітників і розробляти індивідуалізовані стратегії проникнення у корпоративне середовище

Інтелектуальні системи дають змогу зловмисникам імітувати лінгвістичний стиль, лексику та комунікативні патерни окремих працівників або керівного складу, що підвищує правдоподібність фішингових повідомлень, надісланих нібито від імені керівників, фінансових директорів чи ІТ-підрозділів. Автоматизовані інструменти фішингу здатні генерувати масові розсилки тисяч високорелевантних електронних листів, адаптованих під посади, обов’язки й особливості поточних проектів у межах організації [40]. Подібні листи можуть містити прохання щодо надання адміністративного доступу, пересилання конфіденційної інформації або виконання несанкціонованих фінансових операцій під виглядом екстрених доручень в рамках бізнес-процесів.

Технології глибокого підроблення аудіо- та відеоінформації (deepfake) забезпечують зловмисникам інструментарій для створення фальшивих телефонних дзвінків і відеоконференцій із повною імітацією голосу й зовнішності довірених осіб. Маніпулятивний потенціал deepfake-технологій

базується на використанні відкритих записів промов, відео- та аудіоматеріалів, які дозволяють сконструювати цифровий двійник особи для проведення цілеспрямованих соціальних маніпуляцій у корпоративному середовищі.

Подібні атаки формують значний ризик несанкціонованого доступу до критичних ресурсів та компрометації ділової репутації підприємства, оскільки співробітники можуть отримати правдоподібні інструкції, які суттєво відрізняються від стандартних процедур взаємодії.

Застосування автоматизованих фішингових систем на основі ШІ дозволяє створювати масштабовані кампанії, у межах яких формуються тисячі індивідуалізованих електронних повідомлень. Такі листи персоналізуються відповідно до службових функцій та поточних обов'язків окремих співробітників, що підвищує ймовірність успішного впливу на цільову аудиторію. Алгоритмічні механізми обробки внутрішньої структури підприємства, інформації про проєкти та бізнес-процеси дають змогу формувати контекстно релевантні звернення, здатні обґрунтовано ініціювати виконання специфічних дій, серед яких можуть бути: надання доступу до інформаційних систем, передача конфіденційних даних або санкціонування фінансових операцій під виглядом невідкладних виробничих потреб.

Окрему небезпеку становить використання deepfake-технологій для створення аудіо- та відеосинтетичних матеріалів, які можуть застосовуватись у телефонних розмовах чи відеоконференціях для імітації автентичного голосу й вигляду довірених осіб. Опрацювання публічно доступних записів виступів керівного складу надає змогу зловмисникам згенерувати високореалістичні підробки для ведення комунікації з підлеглими чи партнерами. Надзвичайно висока схожість таких синтетичних взаємодій із реальною поведінкою сприяє формуванню ситуацій, коли співробітники можуть отримувати накази чи інструкції, які суперечать стандартним практикам безпеки, однак видаються легітимними в умовах інформаційного впливу.

2.4.2 Шахрайські фіноперації

Штучний інтелект дедалі активніше використовується у фінансових шахрайських схемах, що зумовлює істотну трансформацію традиційних підходів до шахрайства у фінансовому секторі. Зловмисники отримують у розпорядження високотехнологічні засоби для автоматизації, індивідуалізації й масштабування атак на банківські установи, кредитні організації та їхніх клієнтів. Інтелектуальні системи здатні здійснювати глибокий аналіз поведінкових патернів користувачів з метою побудови переконливих сценаріїв шахрайських операцій, що ускладнює їхнє розмежування із легітимною фінансовою діяльністю. Впровадження багаторівневих схем обману поєднує технологічні інструменти і психологічний вплив.

Автоматизовані рішення на базі ШІ дозволяють створювати синтетичні особистості із комплексними фінансовими профілями для відкриття фіктивних рахунків, оформлення кредитів та проведення фінансових операцій. Генеративні алгоритми продукують внутрішньо узгоджені дані, включаючи фотографії, біографії, історії кредитування, цифрові копії документів, які здатні пройти автоматизовані процедури ідентифікації KYC. Використання deepfake-технологій для фальсифікації фото- та відеодокументів значно ускладнює процес верифікації справжності даних.

Технології voice cloning забезпечують зловмисникам можливість відтворювати голоси клієнтів для ініціювання транзакцій через телефонні канали й обхід голосової біометричної аутентифікації.

Для навчання систем достатньо коротких фрагментів аудіо, отриманих із соціальних мереж чи голосових повідомлень. Це дозволяє створювати якісні підробки, які вводять в оману як автоматизовані системи, так і операторів контакт-центрів [41].

Таблиця 2.2

Типи фінансового шахрайства з використанням ШІ

Тип шахрайства	**Технологія ШІ**	**Цільові установи**	**Середні збитки (USD)**	**Складність виявлення**
Synthetic identity	Мультимодальні генеративні моделі	Банки, кредитні установи	Високий	Дуже висока
Voice spoofing	Нейронні вокодери	Call-центри банків	Середній–високий	Висока
Deepfake KYC	Генерація обличчя, відеосинтез	Цифрові банки	Середній	Висока
Automated account takeover	Credential stuffing + ML	Всі фінустанови	Середній	Середня
AI trading manipulation	Підкріплювальне навчання	Біржі, брокери	Дуже високий	Низька

Застосування алгоритмічного моделювання на фінансових ринках із залученням методів машинного навчання суттєво змінює характер маніпуляцій з цінами. Сучасні системи дозволяють автоматизовано ідентифікувати вразливі етапи ринкової динаміки та здійснювати координовані операції для штучного формування цінових трендів (зокрема, схем типу pump-and-dump) з точністю та оперативністю, недосяжними для традиційних методів ручного управління.

Фішингові атаки нового покоління, персоналізовані за допомогою ШІ, ґрунтуються на аналізі соціальних мереж і відкритих джерел для формування індивідуально адаптованих сценаріїв отримання банківських реквізитів. Генеративні моделі здатні створювати комунікаційні повідомлення, які з високим ступенем достовірності відтворюють стиль конкретної фінансової установи, враховуючи особистісні обставини клієнта. Типовими залишаються моделі штучних кризових ситуацій чи фальшивих термінових пропозицій, що стимулюють імпульсивні дії жертви без належної верифікації інформації.

Автоматизація схем легалізації доходів, одержаних злочинним шляхом, реалізується за допомогою алгоритмів оптимізації ланцюгів транзакцій між численними рахунками та юрисдикціями. Такі системи аналізують ризики ідентифікації для кожного кроку і динамічно коригують стратегії, щоб мінімізувати сліди походження коштів. Використання машинного навчання підвищує ефективність сегментації сум, таймінгу операцій та вибору проміжних фінансових інститутів, забезпечуючи високий рівень анонімності руху капіталів.

Технології імітації поведінки легітимних власників платіжних карток за допомогою ШІ відкривають нові перспективи для шахрайства у сфері карткових операцій. Алгоритми формують профілі типових витрат на основі історичних транзакцій, що дозволяє виконувати несанкціоновані операції, залишаючись поза увагою автоматизованих систем моніторингу.

Використання ШІ для корпоративного шахрайства типу Business Email Compromise базується на глибокому аналізі внутрішніх інформаційних потоків організації з метою створення високодостовірних запитів на фінансові перекази від імені керівництва. Технологічне поєднання імітації стилю корпоративного листування та побудови реалістичних ділових сценаріїв забезпечує додаткову переконливість атак, особливо у випадку координації між декількома каналами комунікації.

Автоматизація страхового шахрайства відбувається завдяки генерації синтетичних медичних, технічних або юридичних документів, які слугують підтвердженням неправомірних страхових вимог. Штучний інтелект забезпечує узгодженість і логічну цілісність фальшивих історій нещасних випадків, хвороб чи пошкоджень майна, що значно підвищує ефективність масових шахрайських звернень.

Кіберзлочинці застосовують ШІ для здійснення таргетованого вимагання у фінансових установ із використанням детального аналізу ІТ-інфраструктури для ідентифікації найбільш цінних активів для шифрування. Алгоритмічні моделі дають змогу прогнозувати економічні наслідки інциденту й оптимізувати

стратегії атак, а персоналізовані повідомлення враховують особливості діяльності жертви та її фінансовий потенціал [42].

Впровадження нових моделей моніторингу та контролю у фінансових організаціях стає обов’язковим у зв’язку з адаптивністю алгоритмів шахрайства. Традиційні системи, що працюють за фіксованими правилами, не забезпечують належного захисту перед динамічними ШІ-атаками, а балансування між рівнем безпеки й зручністю для користувачів вимагає інноваційних рішень.

Глобальні проблеми координації протидії ШІ-індукованому фінансовому шахрайству обумовлені нерівномірністю правового регулювання, швидкою еволюцією технологій і різноманіттям юрисдикцій. Недосконалість нормативної бази й наявність міждержавних прогалин створюють сприятливе середовище для транскордонних схем, що ускладнює своєчасний обмін інформацією між банківськими структурами та правоохоронними органами.

2.4.3 Репутаційні збитки та зниження довіри клієнтів

Використання сучасних технологій штучного інтелекту у зловмисних намірах формує тривалу та системну загрозу для організацій, впливаючи як на їхню ринкову позицію, так і на рівень довіри з боку клієнтів і партнерів, що у підсумку може зумовити суттєве послаблення фінансової стійкості суб'єктів господарювання.

Особливу небезпеку становлять дипфейк-атаки, спрямовані на дискредитацію корпоративних лідерів або ключових представників компанії. Слід зазначити, що навіть у разі подальшого спростування й видалення подібного контенту, його початковий вплив на репутацію компанії, з огляду на швидкість поширення інформації в цифрових мережах, може зберігатися протягом тривалого періоду [43].

Крім того, координовані кампанії дискредитації, що ґрунтуються на використанні ботнетів та генеративних моделей штучного інтелекту, надають

змогу масово продукувати й поширювати негативний інформаційний контент у соціальних мережах, на спеціалізованих форумах та платформах для відгуків.

Таблиця 2.3

Типи репутаційних атак з використанням ШІ

Тип атаки	**Технологія штучного інтелекту**	**Цільова платформа**	**Швидкість поширення**	**Потенційні втрати (% ринкової вартості)**
Дипфейк-скандал за участі керівника	Синтез відео	YouTube, X (Twitter)	Дуже висока	5–25%
Масове генерування негативних відгуків	Генерація тексту	Google Reviews, Yelp	Висока	2–10%
Поширення фейкових новин про компанію	Генерація природної мови (NLP)	Новинні онлайн-ресурси	Середня	10–30%
Синтетичні інсайдерські витоки	Генерація документів	WikiLeaks-подібні платформи	Низька	15–40%
AI-генерований контент про продукцію	Мультимодальні AI-системи	Соціальні мережі	Висока	3–15%

Генеративні моделі штучного інтелекту дедалі частіше використовуються для створення переконливих фальшивих внутрішніх документів, які стилістично та структурно імітують автентичну корпоративну документацію, що містить достовірні на вигляд дані про організаційні процеси, структуру підприємства або можливі неетичні практики..

Маніпуляції із продуктово-сервісними рейтингами й оглядами здійснюються шляхом використання ШІ для автоматизованого генерування негативних відгуків, що імітують різноманітні стилі письма та досвіду користування продукцією. Такі системи здатні аналізувати реальні позитивні

відгуки для створення достовірно контрастних негативних рецензій, що ускладнює їх виявлення як підробки [44].

Скоординоване мультиканальне поширення синтетичної інформації сприяє ефекту cascade failure, унаслідок чого швидкість та охоплення фейкових повідомлень можуть перевищувати можливості компанії щодо оперативного реагування та спростування неправдивої інформації.

Секторні атаки із застосуванням ШІ спрямовані на підрив довіри до цілих галузей через створення синтетичних доказів, що стосуються системних проблем, екологічних порушень або ризиків для здоров'я.

Окрему категорію становлять персоналізовані атаки на ключових співробітників, для яких ШІ здійснює аналіз цифрового сліду з метою генерування компрометуючих матеріалів — від підроблених доказів неетичної поведінки або професійних помилок до фабрикації особистих скандалів.

У сукупності фінансові наслідки репутаційних втрат охоплюють зниження ринкової капіталізації, скорочення обсягів реалізації продукції, втрату клієнтської та партнерської бази, а також суттєве зростання витрат на реалізацію антикризових PR-кампаній, юридичне супроводження й технічні заходи з протидії дезінформаційним впливам. Згідно з дослідженнями, компанії, що постраждали від серйозних репутаційних криз, можуть втрачати від 10% до 25% ринкової вартості. Також на Рисунок Б-4 додатку Б було наведено динаміку зростання кількості недостовірних новинних сайтів, згенерованих штучним інтелектом, за даними NewsGuard (2023–2024 роки).

Висновки за розділом 2

У другому розділі було проведено системний аналіз основних ризиків, що виникають у результаті зловмисного використання моделей штучного інтелекту.

Окрема увага приділена загрозам, які формуються внаслідок застосування генеративних систем, великих мовних моделей, дипфейків і мультимодальних

архітектур. Встановлено, що впровадження ШІ сприяє автоматизації кібератак, масштабуванню фішингових кампаній.

Проаналізовано потенціал дипфейків як інструменту впливу на інформаційне середовище, зокрема у сферах шантажу, дискредитації та фальсифікації доказів. Виявлено, що генеративні моделі дедалі частіше використовуються у дезінформаційних кампаніях, що охоплюють соціальні мережі й спрямовані на підрив суспільної стабільності.

Зроблено висновок, що ефективне реагування на зазначені загрози потребує інтегрованих технічних і організаційних рішень, удосконалення систем виявлення та нормативно-правового супроводу в умовах швидкої еволюції технологій ШІ.

РОЗДІЛ 3
СИСТЕМНИЙ ПІДХІД ДО ВИЯВЛЕННЯ ТА ПРОТИДІЇ ЗАГРОЗАМ ШІ В КІБЕРПРОСТОРІ

3.1 Технічні методи виявлення фальшивого контенту

У попередніх розділах були визначені основні проблеми, пов’язані із застосуванням штучного інтелекту: спрощення розробки шкідливого ПЗ, автоматизація кібератак, дезінформація на локальному та глобальному рівнях, а також ризики безвідповідального чи неконтрольованого впровадження ШІ.

Відповідно, у цьому розділі є розглянути підходи та практичні механізми їх подолання для забезпечення безпеки та надійного використання штучного інтелекту.

Аналіз метаданих відіграє значну роль у процесі ідентифікації штучно згенерованого контенту шляхом дослідження технічних атрибутів цифрових файлів, що супроводжують основний інформаційний масив. Метадані містять детальні відомості про способи створення файлу, характеристики використаного обладнання, тип програмного забезпечення, часові параметри та просторові координати [45]. Глибока експертиза таких параметрів дозволяє встановлювати невідповідності та аномалії, котрі можуть слугувати індикаторами штучного генезису інформації або свідчити про втручання у структуру файлу після його початкового формування.

Цифрові сліди, що містяться у метаданих графічних файлів, відображають широкий спектр характеристик – від моделі пристрою захоплення, параметрів експозиції, ISO та діафрагми, до специфіки налаштувань зйомки. Вироблені генеративними алгоритмами зображення часто характеризуються нестачею достовірності або узгодженості таких параметрів, що призводить до появи неочевидних суперечностей між технічними атрибутами та реальними фізичними властивостями файлу. Проведення детального аналізу EXIF-інформації дозволяє виявляти випадки застосування стороннього програмного

забезпечення для генерації або маніпуляцій із зображенням, а також атипові комбінації технічних налаштувань, що виходять за межі стандартної логіки процесу фотозйомки. Також на (Рисунку Б-7) додатку Б було наведено зразок архітектури системи виявлення синтетичного контенту.

Окрему аналітичну цінність становлять часові маркери, присутні у структурі метаданих. Вони здатні сигналізувати про факт синтетичного походження файлів або подальших модифікацій контенту. Для штучно створених зображень властиві невідповідності між різними категоріями часових записів, відсутність хронологічної послідовності у файловій системі чи аномально високі темпи формування великої кількості контенту. Поглиблений порівняльний аналіз міток створення, редагування й доступу до об’єкта дає змогу ідентифікувати типові шаблони, притаманні автоматизованим системам генерації цифрових файлів [46].

Таблиця 3.1

Характеристики метаданих для різних типів контенту

Тип файлу	**Типи метаданих**	**Індикатори синтетичного походження**	**Імовірність коректної ідентифікації**	**Стійкість до підробки**
MP4 відео	Кодек, fps, роздільність	Нетипова структура кодування	80-90%	Низька
JPEG зображення	EXIF, JFIF, Adobe	Аномалії технічних параметрів пристрою	70-75%	Середня
WAV аудіо	Частота дискретизації, біт-рейт	Штучні характеристики шуму	70-80%	Висока
PDF документи	Creator, Producer, timestamps	Автоматизовані інструменти створення	85-95%	Низька
PNG графіка	Палітра, стиснення	Генеративні артефакти	65-75%	Висока

3.2 Рекомендації зі створення ПЗ для захисту на державному рівні

З урахуванням загроз, висвітлених у попередніх розділах, доцільною видається ініціатива щодо створення інтегрованої комплексної системи виявлення та реагування на атаки із застосуванням штучного інтелекту (КСЗШІ), адаптованої до державного та критичного корпоративного сегменту. Архітектурно така система має поєднувати принципи SIEM (Security Information and Event Management), DLP (Data Loss Prevention), та сучасні алгоритми ML/LLM-аналізу загроз в одному функціональному середовищі. На (рисунку 3.1) наведено зразок архітектури мережі, що може підпадати під таку атаку.

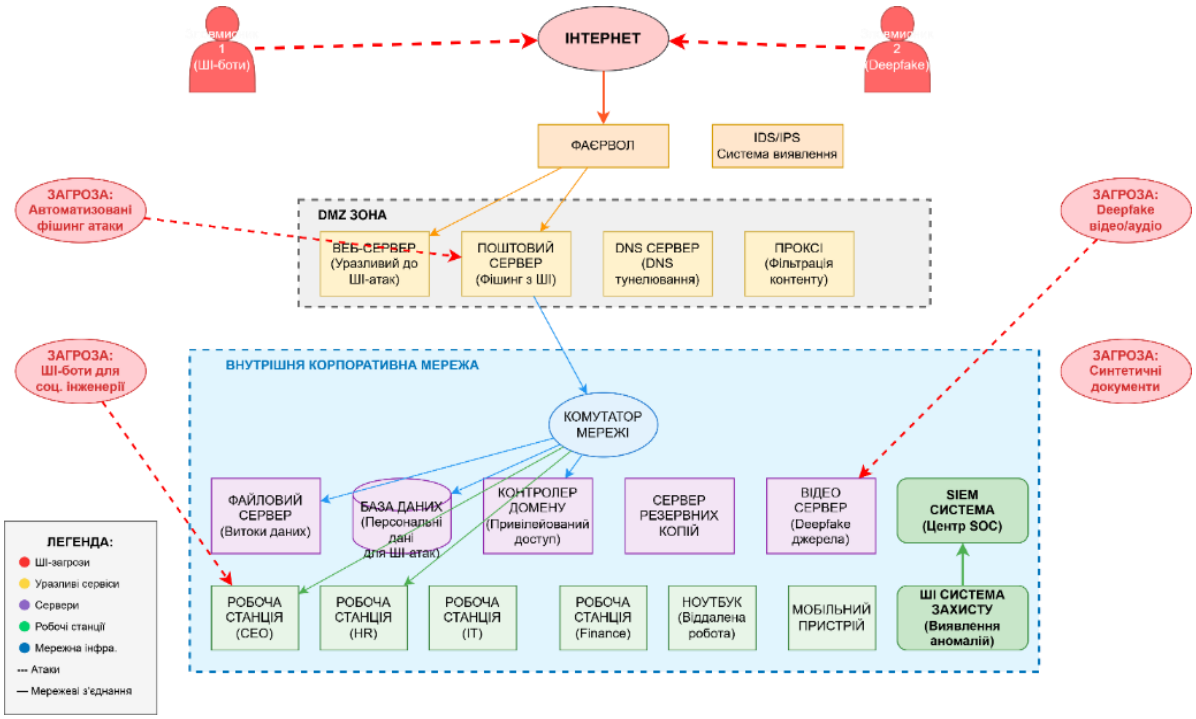

Рисунок 3.1 – Архітектура корпоративної мережі під загрозою ШІ-атак

Модуль протидії фішингу

Першим і одним із центральних елементів КСЗШІ є інтелектуальний модуль фільтрації фішингових повідомлень. Його функціональність має виходити за межі класичних сигнатурних методів і включати багаторівневу мовно-семантичну експертизу вхідного трафіку.

Особливої уваги потребують повідомлення, згенеровані за допомогою великих мовних моделей, таких як GPT-4 або Claude 3, оскільки вони здатні адаптуватися під стилістику офіційного листування, виявляти слабкі місця у психолінгвістичному профілі одержувача, формувати соціально-інженерні шаблони, що обходять класичні системи фільтрації. Для боротьби з такими ризиками особливої уваги заслуговує використання федеративного навчання (Federated Learning, FL) у поєднанні з трансформерними моделями (на кшталт BERT) для побудови динамічних антифішингових систем, що можуть функціонувати в умовах організаційної конфіденційності без розкриття приватної електронної кореспонденції.

У рамках такої системи захисту, фільтрація фішингових повідомлень із генеративними ознаками має спиратися на багаторівневий аналіз — на рівні лексико-семантичної структури, елементів стилю, заголовків та вмісту повідомлення. У порівнянні з класичними методами, що спираються на ручне виділення ознак, трансформерні моделі демонструють здатність до глибинного контекстного розуміння, що є вирішальним у виявленні мовних аномалій, згенерованих ШІ (наприклад, великі мовні моделі, такі як GPT-4, мають тенденцію до надмірної формалізації мови та специфічної структури речень).

У дослідженні "Evaluation of Federated Learning in Phishing Email Detection" від 2020 року [47] запропоновано об'єднати модель BERT із підходом федеративного навчання, що дозволяє досягти практично ідентичної точності виявлення фішингу у порівнянні з централізованим навчанням (наприклад, 96.1% проти 96.2% при 5 клієнтах). Особливе значення має той факт, що FL дозволяє зберігати конфіденційність даних — модель навчається локально, а узагальнення знань відбувається через агрегацію ваг моделей, а не даних.

Для практичного впровадження такої системи на державному або корпоративному рівні доцільно передбачити гібридну архітектуру КСЗШІ, в якій модуль фільтрації фішингу включатиме наступне:

- Ініціалізація моделі BERT, донавченої на локальному наборі корпоративної пошти, в якій передбачено структури фішингових атак з

ознаками LLM (наприклад, маскування під технічну підтримку, зловживання термінологією, відсутність персоналізації).

- Формалізація генеративних шаблонів**,** включаючи синтаксичні патерни та стилістичні маркери (надмірне використання модальних дієслів, логічних конекторів, структурованих форматів часу).
- Впровадження федеративного оновлення з можливістю розгортання центрального вузла при Держспецзв'язку або СБУ, який синхронізує моделі клієнтів без доступу до даних.
- Інтеграція з SPF/DKIM/DMARC модулями перевірки заголовків та верифікації маршрутів повідомлення для зменшення кількості помилкових позитивних спрацювань.
- Аналіз фішингових стратегій, що використовують дипфейкові вкладення або голосові повідомлення, які маскують соціальні атаки. Для цього можлива інтеграція з детекторами синтетичного контенту, зокрема мовних чи аудіоаналізаторів, які на основі глибинних спектральних ознак виявляють неприродні параметри голосу або шумових артефактів, характерних для моделей на кшталт VALL-E чи Voicebox.

Також у додатку А на (Таблиці А-3) було наведено більш конкретне порівняння порівняння ефективності методів виявлення фішингових повідомлень, згенерованих ШІ та було наведено приклади ключових ознак фішингових повідомлень, згенерованих ШІ.

Модуль протидії графічним дипфейкам

Наступною критичною підсистемою є блок виявлення мультимедійного контенту, створеного або модифікованого за допомогою генеративних моделей ШІ. Актуальність даного компоненту обумовлена широким застосуванням дипфейків у дискредитаційних кампаніях, атаках із застосуванням соціальної інженерії, та навіть в операціях шантажу.

Алгоритмічна реалізація повинна базуватися на згорткових нейронних мережах із вбудованими механізмами виявлення артефактів компресії, спектральної модуляції, відсутності мікровиразів обличчя (для відео), а також

фазової дисперсії в аудіосигналі. Для верифікації автентичності зображень і відео доцільним є використання моделей, натренованих на відкритих джерелах.

До прикладу, в дослідженні DeepFakes Detection: the DeeperForensics Dataset and Challenge запропоновано системний підхід до валідації алгоритмів детекції дипфейків у складних реальних умовах, що враховують спотворення, типові для цифрових комунікацій (розмитість, стиснення, освітлення, цифрові шуми). Зокрема, представлена база даних включає понад 60 тисяч відеофрагментів, модифікованих із використанням 6 різних генеративних моделей (DF-VAE, FaceSwap, Face2Face, NeuralTextures тощо), які були додатково змодельовані у 35 реалістичних умовах.

На основі цього корпусу розроблено архітектуру детектора, побудовану на згорткових нейронних мережах з попередньо навченими фільтрами, адаптованими до артефактів на рівні пікселів. Найефективніші результати показала модель XceptionNet, модифікована для виявлення локальних просторових аномалій у зонах обличчя та шкіри. Структуру цього детектора наведено на (рисунку 3.2).

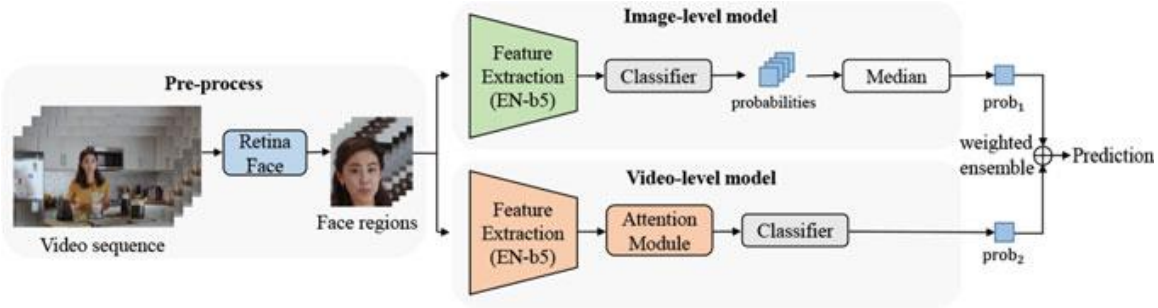

Рисунок 3.2 – Структурне відображення процесу виявлення DeepFake-відео

Аналіз здійснюється у два етапи: спочатку обличчя автоматично виділяються з відео (детектор RetinaFace), далі окремі кадри проходять ознаковий аналіз і класифікацію (EN-b5), а медіана ймовірностей визначає оцінку для зображень (prob1). Паралельно з цього ж набору зображень формується послідовність для відеорівневої моделі з модулем уваги, що фокусує класифікатор на ключових кадрах (prob2). Фінальний результат отримують

шляхом зваженого об'єднання prob1 і prob2, що підвищує точність виявлення DeepFake.

Результати тестування підтвердили, що найбільшу складність для алгоритмів становлять дипфейки, оброблені в умовах реального світу, зокрема при зміні масштабу, накладанні фільтрів чи повторному кодуванні. Це обумовлює необхідність використання не лише чистих датасетів, а й моделювання умов, наближених до бойових, у процесі навчання системи.

Модуль протидії аудіодипфейкам

В аудіосфері варто інтегрувати архітектури типу Wav2Vec 2.0, ECAPA-TDNN, що здатні виявляти синтетичний тембр, нехарактерні спектрограми та часові розбіжності, які не зустрічаються у природному мовленні людини.

У дослідженні під назвою "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation" запропоновано підхід до детекції дипфейків у сфері автоматичної голосової верифікації, що базується на самонавчанні (self-supervised learning) та застосуванні моделі wav2vec 2.0 у комбінації з AASIST (Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Network) [48]. Ключова перевага цієї архітектури полягає у здатності узагальнювати характеристики справжнього та фальшивого мовлення навіть при суттєвих варіаціях даних.

Фреймворк побудований на основі моделі wav2vec 2.0 XLS-R, яка проходить попереднє навчання на багатомовних корпусах із реального мовлення (CommonVoice, MLS, VoxPopuli, VoxLingua107). Після цього здійснюється донавчання на спеціалізованих наборах даних ASVspoof із фейковими голосами. Завдяки цій двофазній схемі модель демонструє стійкість до атак, згенерованих понад 100 різними алгоритмами. Для виявлення дипфейків у результаті аналізу аудіосигнал трансформується у спектро-часову репрезентацію, що далі обробляється графовими модулями, зображену на (рисунку 3.3).

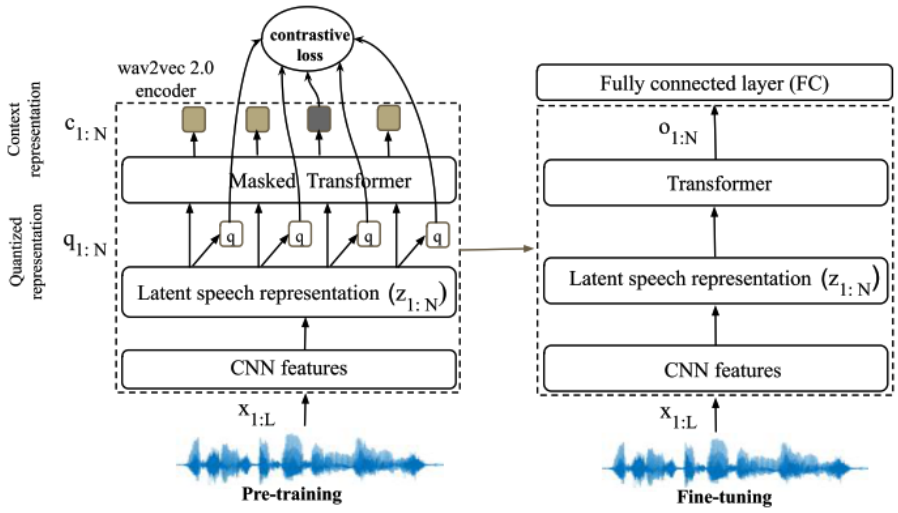

Рисунок 3.3. - Структура попереднього навчання wav2vec 2.0

На етапі pre-training сирий аудіосигнал проходить через згорткову нейромережу (CNN), далі формується латентне представлення мовлення, яке частково маскується та подається у трансформер. Одночасно певні латентні вектори квантуються. Для навчання використовується контрастивна функція втрат, що дозволяє моделі навчитися виділяти змістовні ознаки без використання анотованих даних.

На етапі fine-tuning модель донавчається для конкретного завдання: латентні ознаки, отримані з CNN і трансформера, проходять через повнозв'язний шар для класифікації справжнього або підробленого мовлення. Весь процес забезпечує формування стійких та інформативних представлень, що підвищує якість виявлення різних типів спуфінгу та deepfake-атак навіть за умови обмежених розмічених даних

У процесі навчання система додатково підсилюється методом Data Augmentation на основі RawBoost, який моделює реалістичні умови компресії, завад і шумів. Це дозволяє моделі не лише ефективно виявляти дипфейки у лабораторному середовищі, а й демонструвати високу стійкість до атак у динамічному середовищі (наприклад, при передачі через VoIP або месенджери).

Результати експериментальної перевірки підтверджують ефективність підходу: при поєднанні wav2vec 2.0, self-attentive aggregation та DA було досягнуто найнижчого серед відомих рівня помилок EER на ASVspoof 2021 (LA: 0.82%, DF: 2.85%), що становить покращення на понад 85% у порівнянні з базовими методами.

У контексті впровадження на рівні підприємства чи урядових структур доцільно інтегрувати систему як back-end для внутрішніх поштових серверів, VoIP шлюзів або SIEM-платформ.

Для аналізу PDF, DOCX та інших офісних форматів необхідно запровадити мультифазну обробку: первинне сигнатурне сканування, аналіз структури документа на предмет прихованих скриптів, а також динамічне виконання у віртуальному середовищі (sandbox) з аналізом поведінкових характеристик об'єкта.

Не менш важливою є реалізація поведінкової аналітики на рівні мережевого моніторингу. Цей блок має використовувати несупервізоване навчання для виявлення аномалій у трафіку, які можуть вказувати на вторгнення, здійснене як людиною, так і автономним агентом ШІ. Тут застосовуються алгоритми типу Isolation Forest, One-Class SVM, а також AutoEncoder-мережі, які дозволяють виявляти аномальні сесії доступу, неприродні патерни звернень до API, або надмірну швидкість сканування портів. Аналіз повинен здійснюватися в реальному часі із підтримкою потокової обробки (наприклад, Apache Flink або Kafka Streams). На (рисунку Б-8) додатку Б було наведено візуалізацію гіперплощини One-Class SVM та кластеризації нормальних і аномальних точок.

Іншою функціональною одиницею має стати підсистема оцінки автентичності контенту з месенджерів — зокрема відео- й аудіоповідомлень. Її архітектура має поєднувати гібридний підхід: верифікацію цифрового підпису (у випадку використання протоколів end-to-end encryption), визначення наявності водяного знаку, а також нейромережевий аналіз вмісту. Доцільним є використання комбінованого пайплайну, в якому перший рівень визначає формат повідомлення, другий — проводить попередню декомпозицію, а третій

— здійснює аналіз через навчений класифікатор. Такі підсистеми здатні виявити, наприклад, синтетичні відео, де рух губ не синхронізований з аудіо, або аудіофайли, згенеровані голосовими клонами.

3.3 Організаційно-процедурні заходи з протидії ризиків ШІ

Політики інформаційної безпеки, як сукупність офіційно затверджених документів організації, регламентують принципи, процедури та вимоги до захисту інформаційних ресурсів від загроз, що виникають у зв'язку із застосуванням штучного інтелекту для здійснення зловмисних дій. У таких політиках мають бути враховані всі етапи обігу інформації: починаючи від її генерування, зберігання, передачі, обробки і закінчуючи знищенням, із визначенням вимог до контролю доступу, моніторингу та реагування на інциденти.

Зважаючи на сучасну динаміку розвитку кіберзагроз, пов'язаних із використанням технологій ШІ, доцільним є перегляд та актуалізація існуючих політик інформаційної безпеки шляхом впровадження нових положень щодо виявлення, аналізу та протидії синтетичному контенту, автоматизованим атакам і іншим різновидам інтелектуальних загроз. При цьому особливу увагу слід приділяти розробці механізмів управління ризиками, пов’язаними з впровадженням та експлуатацією систем штучного інтелекту в організаційному середовищі [50].

Структурні елементи політики інформаційної безпеки мають містити детально прописані ролі та функціональні обов’язки працівників з урахуванням аспектів протидії загрозам, що породжуються використанням технологій штучного інтелекту.

Технічні вимоги, визначені у політиках, повинні фіксувати правила застосування систем для виявлення штучно синтезованого контенту, автентифікації цифрових медіаданих та верифікації джерел інформації. Регламентація має містити мінімально допустимі характеристики програмних і

апаратних засобів захисту, а також вимогу регулярної модернізації захисних рішень відповідно до еволюції актуальних загроз та появи нових інструментів атаки.

Процедури класифікації інформаційних активів повинні передбачати оцінку специфічних ризиків, які виникають унаслідок імовірного використання даних організації для генерування синтетичного контенту. Необхідно впроваджувати підвищені стандарти захисту для тих категорій інформації, що можуть бути об’єктом маніпуляцій або використані у векторі атаки, зокрема для персональних зображень співробітників, корпоративної документації, аудіо- чи відеофрагментів тощо.

Таблиця 3.2

Ключові елементи політики інформаційної безпеки у протидії загрозам ШІ

Компонент політики	**Галузь застосування**	**Періодичність перегляду**	**Ступінь деталізації**	**Відповідальні підрозділи**
Технічні регламенти	Інформаційна інфраструктура	Щоквартально	Високий	IT-директор, CISO
Алгоритми реагування	Управління інцидентами	Один раз на рік	Середній	Група безпеки
Освітні ініціативи	Весь штат організації	Раз на півроку	Низький	Відділ кадрів, служба захисту
Процедури аудиту й контролю	Всі бізнес-процеси	Щорічно	Високий	Внутрішній аудит
Вимоги нормативної відповідності	Дотримання регуляторних норм	За необхідністю	Дуже високий	Юридичний відділ, служба compliance

Політики управління доступом до інформаційних ресурсів повинні передбачати чіткі нормативи застосування ШІ-систем під час обробки

конфіденційних корпоративних даних. Необхідно встановити суворі обмеження щодо автоматизованого опрацювання чутливої інформації, зокрема визначити правила використання хмарних сервісів на основі ШІ, заборонити несанкціоноване завантаження корпоративних даних у відкриті платформи машинного навчання та забезпечити вимоги до локального зберігання критично важливої інформації. Окремо слід визначити порядок застосування персональних цифрових асистентів працівниками в рамках виконання службових обов'язків.

Аудит та моніторинг систем інформаційної безпеки, в особливості КСЗШІ має здійснюватися на регулярній основі з метою оцінювання ефективності заходів протидії загрозам, що походять від застосування ШІ. Політики повинні визначати набір релевантних метрик, що дозволяють оцінити ступінь захищеності, а також регламентувати проведення тестування систем виявлення синтетичного контенту та аналізу тенденцій у сфері кібербезпеки. Отримані результати аудитів мають стати підґрунтям для удосконалення чинних політик та процедур.

Система управління постачальниками та контрагентами повинна включати механізми верифікації їх відповідності стандартам інформаційної безпеки та готовності до протидії ШІ-загрозам. Політики організації мають визначати перелік вимог до зовнішніх партнерів, регламентувати процедури верифікації цифрових комунікацій, а також передбачати наявність ефективних засобів контролю доступу до спільних інформаційних ресурсів. Особливу увагу варто приділити кіберзахисту ланцюгів постачання, які залишаються особливо уразливими до маніпуляцій через синтетичний контент [51].

Підтримка актуальності політик інформаційної безпеки повинна ґрунтуватися на систематичному аналізі нових кіберзагроз, технологічних інновацій та досвіду реагування на попередні інциденти. Для забезпечення гнучкості організації доцільно створити експертні робочі групи, які будуть відповідальні за регулярне оновлення політик, перегляд процедур та організацію навчання персоналу у відповідності до новітніх тенденцій.

3.4 Регуляторно-правові механізми

Національні ініціативи у сфері регулювання ризиків, пов'язаних із використанням штучного інтелекту, являють собою сукупність нормативно-правових, організаційно-адміністративних та програмних заходів, спрямованих на формування цілісної правової системи, здатної ефективно протидіяти зловмисному застосуванню ШІ-технологій і забезпечувати сталий розвиток цифрової інфраструктури. Відповідна політика має передбачати не лише мінімізацію специфічних ризиків для національної кібербезпеки, а й створення сприятливого середовища для інновацій, включаючи механізми державної підтримки технологічного бізнесу та наукових розробок.

Конституційні основи регулювання у сфері протидії ШІ-загрозам базуються на принципах забезпечення прав на захист персональних даних, охорону приватного життя, свободу доступу до інформації та гарантування національної безпеки. Зазначені положення формують фундамент для розробки галузевого та спеціалізованого законодавства, спрямованого на протидію кіберзагрозам у цифровому просторі [52].

Спеціальні законодавчі акти у сфері кібербезпеки мають включати нормативні визначення поняття штучного інтелекту та його зловмисного застосування, встановлення кримінальної й адміністративної відповідальності за розробку, поширення й використання шкідливих ШІ-інструментів, а також механізми координації між державними органами у реагуванні на інциденти. Окремо слід передбачити положення щодо регулювання технологій deepfake, автоматизованих систем кібератак та інших сучасних інструментів інформаційного впливу.

Таблиця 3.3

Типи національних правових актів для регулювання ШІ-загроз

Тип акту	**Сфера регулювання**	**Механізми впливу**	**Відповідальність**	**Термін дії**
Конституційні норми	Основні права	Судовий контроль	Конституційний суд	Безстроково

Продовження таблиці 3.3

Кодекси	Кримінальна/адмін відповідальність	Санкції, покарання	Суди	Без обмежень
Галузеві закони	Кібербезпека, ШІ	Ліцензування, контроль	Регулятори	До скасування
Нормативно-правові акти підзаконного рівня	Технічні стандарти	Сертифікація	Галузеві органи	3-5 років
Стратегії	Політичні напрями	Планування, фінансування	Уряд	5-10 років

3.5 Міжнародне співробітництво у сфері нейтралізації загроз ШІ

Міжнародне співробітництво у сфері нейтралізації загроз, пов'язаних із використанням технологій штучного інтелекту, формується як багаторівнева система організованої взаємодії між державними структурами, наднаціональними інституціями та недержавними суб'єктами з метою координації заходів щодо запобігання та протидії транснаціональним кіберінцидентам. Оскільки інформаційний простір має глобальний характер і не обмежується національними кордонами, ефективна протидія ШІ-загрозам вимагає скоординованого залучення міжнародної спільноти [53].

Механізми багатостороннього міжнародного правового регулювання передбачають створення договірних основ для міждержавної взаємодії у сфері кібербезпеки, з інтеграцією аспектів, пов’язаних із використанням штучного інтелекту у протиправній діяльності. Існуючі конвенції, зокрема Будапештська конвенція Ради Європи про кіберзлочинність, потребують адаптації та розширення сфери дії з урахуванням появи нових категорій правопорушень, що виникають внаслідок впровадження генеративних та автономних ШІ-систем. Новітні багатосторонні угоди можуть передбачати впровадження спеціалізованих органів для координації розслідувань, розробку єдиних стандартів фіксації, збереження та обміну доказовою інформацією, а також регулярний перегляд та вдосконалення механізмів взаємодії.

У межах двосторонньої співпраці між окремими державами доцільно створювати спеціальні комунікаційні канали між правоохоронними структурами, розробляти спільні процедури оперативного обміну інформацією про ШІ-інциденти та організовувати координаційні заходи щодо протидії високотехнологічним загрозам. Динамічність таких форматів сприяє адаптації міжнародного співробітництва до специфічних потреб окремих країн.

Таблиця 3.4

Механізми міжнародного співробітництва проти ШІ-загроз

Механізм	**Учасники**	**Сфера дії**	**Правовий статус**	**Ефективність**
ООН комітети	Держави-члени	Глобальна	Рекомендаційний	Низька
Регіональні організації	Країни регіону	Регіональна	Обов'язковий	Середня
Двосторонні угоди	2 держави	Спеціалізована	Обов'язковий	Висока
Техн. консорціуми	Компанії, держави	Стандарти	Добровільний	Середня
Правоохоронні мережі	Поліція, спецслужби	Оперативна	Обмежений	Висока

Регіональні організації забезпечують проміжний рівень співробітництва між глобальними багатосторонніми ініціативами та індивідуальними двосторонніми домовленостями, сприяючи формуванню єдиних стандартів реагування на загрози, що виникають внаслідок застосування штучного інтелекту. Такі інтеграційні об’єднання, як Європейський Союз або АСЕАН, мають можливість розробляти уніфіковані процедури і протоколи для забезпечення адекватної протидії ШІ-загрозам, враховуючи при цьому регіональні соціокультурні та правові особливості. Регіональний вимір співробітництва підвищує ефективність втілення превентивних та реактивних заходів, підтримуючи оптимальний баланс між локалізацією та масштабністю реагування.

Технічне співробітництво на міжнародному рівні полягає у спільній розробці технологічних стандартів безпеки, взаємному обміні інноваційними технічними рішеннями та координації науково-дослідної діяльності, спрямованої на розробку засобів детекції та нейтралізації ШІ-загроз. У цьому контексті особливу роль відіграють комплексні системи захисту інформації (КСЗШІ), які стають основою для імплементації новітніх стандартів безпеки у державному та корпоративному секторі.

Оперативна взаємодія між правоохоронними структурами різних країн передбачає створення спеціалізованих каналів для екстреного обміну інформацією про кіберінциденти, пов’язані із застосуванням ШІ, а також організацію скоординованих розслідувань. Міжнародні агенції, такі як Інтерпол та Європол, формують профільні підрозділи для боротьби з технологічною злочинністю, яка здійснюється із застосуванням інтелектуальних систем.

Інформаційний обмін загрозами, що стосуються ШІ, потребує впровадження захищених протоколів та механізмів обмеженого доступу для захисту чутливої інформації. Автоматизовані платформи обміну індикаторами компрометації та ознаками атак (IoC) у режимі реального часу, а також сучасні КСЗШІ, дозволяють забезпечити своєчасну ідентифікацію нових векторів загроз, а ефективна класифікація й маркування даних захищає національні інтереси.

Дипломатичні механізми взаємодії охоплюють вироблення міжнародних норм поведінки у кіберпросторі та розробку процедур деескалації конфліктів, пов’язаних із застосуванням інноваційних технологій. Глобальні угоди можуть встановлювати обмеження на створення, тестування чи застосування певних категорій автономних чи деструктивних ШІ-систем.

Міжнародне приватно-державне партнерство сприяє залученню технологічних корпорацій до процесу виявлення, ідентифікації й усунення ШІ-загроз. Платформи співпраці можуть використовуватись для спільного ведення баз знань про новітні методи генерації та розповсюдження синтетичного контенту, а також для реалізації заходів із його швидкої ідентифікації та

блокування. Корпоративні стандарти відповідальності стимулюють компанії до активної участі у формуванні кіберстійкості цифрового простору.

Моніторинг ефективності міжнародного співробітництва передбачає застосування систематичних критеріїв та індикаторів для оцінки дієвості реалізованих заходів. Аналіз результатів спільних проектів та ініціатив дозволяє ідентифікувати найуспішніші практики й виявляти напрямки для подальшого вдосконалення механізмів взаємодії.

Таблиця 3.5

Ефективність різних форм міжнародного співробітництва

Форма співробітництва	**Швидкість реагування**	**Сфера впливу загроз**	**Політична підтримка**	**Ресурсні вимоги**
Цілодобова оперативна координація	Дуже висока	Спеціалізована	Висока	Середні
Технічне	Середня	Мультисекторальна	Середня	Високі
Дипломатичний діалог	Низька	Глобальна	Дуже висока	Низькі
Правове	Низька	Регламентована	Висока	Середні
Економічне	Середня	Розширена	Середня	Дуже високі

Вагомими чинниками, що істотно ускладнюють формування дієвої системи міжнародної співпраці у сфері протидії загрозам штучного інтелекту, виступають глибокі відмінності у національних нормативно-правових підходах до кібербезпеки, протистояння геополітичних та економічних інтересів, а також нерівномірність розвитку захисної інфраструктури у різних державах. Держава, зберігаючи пріоритет суверенітету над інформаційними ресурсами та критичними комунікаційними системами, часто обмежує можливості для інтеграції національних кіберзахисних сегментів у глобальні механізми, ускладнюючи формування міждержавних процедур раннього попередження й реагування на загрози.

Сучасна міжнародна система кібербезпеки спрямована на створення глобальної архітектури протидії ШІ-загрозам шляхом багаторівневої

інституціоналізації співпраці та впровадження інноваційних технологій для захищеного інформаційного обміну. Посилення ролі спеціалізованих міжнародних організацій сприяє гармонізації стандартів, уніфікації процедур і забезпечує координацію стратегій реагування на загрози, що виникають внаслідок розвитку ШІ.

ВИСНОВКИ

У рамках кваліфікаційної роботи було досліджено актуальні ризики зловмисного застосування генеративних моделей штучного інтелекту, проаналізовано механізми ескалації кіберзагроз за рахунок автоматизації атак, фішингових кампаній, дезінформаційних операцій і створення deepfake-контенту, оцінено деструктивний потенціал мультимодальних систем та технологій синтетичних медіа у корпоративному й державному секторах, визначено специфіку загроз цілісності, конфіденційності та достовірності даних, а також розроблено комплекс рекомендацій щодо інтеграції технічних, організаційно-процедурних і регуляторно-правових заходів нейтралізації зазначених ризиків.

Отримані результати засвідчили, що поширення великих мовних моделей (GPT, PaLM, Claude, Gemini) і дифузійних генераторів підсилює емерджентні вектори кібератак, роблячи їх масштабнішими та малопомітними для класичних засобів захисту. Феномен дипфейку, підтверджений практичними кейсами компрометації корпоративної та державної комунікації, доводить можливість цілеспрямованого підриву цифрової автентичності й соціальної довіри через високореалістичні аудіо-, відео- та графічні підробки. Доведено, що найбільш уразливими залишаються вузли внутрішніх інформаційних потоків: поштові шлюзи, системи управління документообігом та сервіси віддаленого доступу, які стають точками входу для ШІ-орієнтованих атак із використанням соціальної інженерії.

Запропонований механізм оцінювання ризиків, що поєднує такі засоби, як поведінкові моделі One-Class SVM, глибокі згорткові мережі для детекції синтетичного контенту та федеративні підходи в антифішингових системах, забезпечує раннє виявлення аномалій без порушення конфіденційності корпоративних даних.

Комплексний підхід, розроблений у роботі, формує методологічне підґрунтя для побудови багаторівневої системи кіберзахисту, здатної адаптуватися до динаміки загроз штучного інтелекту, сприяти вдосконаленню політик інформаційної безпеки, підтримувати державні ініціативи з міжвідомчої координації та забезпечувати науковий базис подальших досліджень у сфері протидії ШІ-загрозам.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. IBM. Machine Learning. [Електронний ресурс] – Режим доступу до ресурсу: https://www.ibm.com/think/topics/machine-learning – Дата звернення: 03.05.2025.

2. Gary Adcock. What Is AI Video Compression? // massive.io, 5 January 2023. [Електронний ресурс] – Режим доступу до ресурсу: https://massive.io – Дата звернення: 03.05.2025.

3. Buhmann, J.; Kuhnel, H. Unsupervised and supervised data clustering with competitive neural networks // [Proceedings 1992] IJCNN International Joint Conference on Neural Networks. Vol. 4. IEEE, 1992. – С. 796–801. doi:10.1109/ijcnn.1992.227220.

4. S. Geman, E. Bienenstock, R. Doursat. Neural networks and the bias/variance dilemma // Neural Computation. – 1992. – Vol. 4, №1. – P. 1–58.

5. ZDNet. What is ChatGPT: how the world's most popular AI chatbot can benefit you? [Електронний ресурс] – Режим доступу до ресурсу: https://www.zdnet.com/article/what-is-chatgpt-how-the-worlds-most-popular-ai-chatbot-can-benefit-you/ – Дата звернення: 03.05.2025.

6. Fung S., Lu X., Zhang C., Li C.-T. DeepfakeUCL: Deepfake Detection via Unsupervised Contrastive Learning.

7. Russell S.J., Norvig P. Artificial Intelligence: A Modern Approach. – Third ed. – Upper Saddle River, New Jersey: Prentice Hall, 2010. – 830, 831 с. – ISBN 978-0-13-604259-4.

8. TeamAI. Understanding Different ChatGPT Models. [Електронний ресурс] – Режим доступу до ресурсу: https://teamai.com/blog/large-language-models-llms/understanding-different-chatgpt-models/ – Дата звернення: 04.05.2025.

9. Wired. OpenAI Sora: Generative AI Video. [Електронний ресурс] – Режим доступу до ресурсу: https://www.wired.com/story/openai-sora-generative-ai-video/ – Дата звернення: 04.05.2025.

10. Peebles W., Xie S. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023. – P. 4195–4205.

11. OpenAI. Microsoft invests in and partners with OpenAI. [Електронний ресурс] – Режим доступу до ресурсу: https://openai.com/index/microsoft-invests-in-and-partners-with-openai/ – Дата звернення: 04.05.2025.

12. Microsoft Bing Blog. Building the New Bing. [Електронний ресурс] – Режим доступу до ресурсу: https://blogs.bing.com/search-quality-insights/february-2023/Building-the-New-Bing – Дата звернення: 05.05.2025.

13. Kotaku. Microsoft Bing AI Image Art: Kirby, Mario. [Електронний ресурс] – Режим доступу до ресурсу: https://kotaku.com/microsoft-bing-ai-image-art-kirby-mario-9-11-nintendo-1850899895 – Дата звернення: 05.05.2025.

14. Double Pulsar. Microsoft Recall on Copilot PC: Testing the Security and Privacy Implications. [Електронний ресурс] – Режим доступу до ресурсу: https://doublepulsar.com/microsoft-recall-on-copilot-pc-testing-the-security-and-privacy-implications-ddb296093b6c – Дата звернення: 05.05.2025.

15. Wired. Google’s Open Source AI TensorFlow Signals Fast-Changing Hardware World. [Електронний ресурс] – Режим доступу до ресурсу: https://www.wired.com/2015/11/googles-open-source-ai-tensorflow-signals-fast-changing-hardware-world/ – Дата звернення: 06.05.2025.

16. Dean, Jeff; Monga, Rajat; et al. TensorFlow: Large-scale machine learning on heterogeneous systems. Google Research. 2015. [Електронний ресурс]. – Режим доступу до ресурсу: https://www.tensorflow.org – Дата звернення: 06.05.2025.

17. DeepMind. Veo. [Електронний ресурс] – Режим доступу до ресурсу: https://deepmind.google/models/veo – Дата звернення: 06.05.2025.

18. Ваврик Ю., Опірський І. ШТУЧНИЙ ІНТЕЛЕКТ: КІБЕРБЕЗПЕКА НОВОГО ПОКОЛІННЯ. Ukrainian Scientific Journal of Information Security. 2024. Vol. 30, № 2. С. 242-254. URL: https://jrnl.nau.edu.ua/index.php/Infosecurity/article/view/19235

19. Respeecher. The Rise of Ethical Voice Cloning in the Deepfake Voice Wars. [Електронний ресурс] – Режим доступу до ресурсу:

https://www.respeecher.com/blog/the-rise-of-ethical-voice-cloning-in-the-deepfake-voice-wars – Дата звернення: 07.05.2025.

20. Tech.co. What is Claude AI? [Електронний ресурс] – Режим доступу до ресурсу: https://tech.co/news/what-is-claude-ai-anthropic – Дата звернення: 07.05.2025.

21. GeeksforGeeks. Generative Adversarial Network (GAN). [Електронний ресурс]. – Режим доступу до ресурсу: https://www.geeksforgeeks.org/generative-adversarial-network-gan/ – Дата звернення: 07.05.2025.

22. TowardsAI. Diffusion Models vs GANs vs VAEs: Comparison of Deep Generative Models. [Електронний ресурс] – Режим доступу до ресурсу: https://pub.towardsai.net/diffusion-models-vs-gans-vs-vaes-comparison-of-deep-generative-models-67ab93e0d9ae – Дата звернення: 08.05.2025.

23. TechTarget. Generative models: VAEs, GANs, diffusion, transformers, NeRFs. [Електронний ресурс] – Режим доступу до ресурсу: https://www.techtarget.com/searchenterpriseai/tip/Generative-models-VAEs-GANs-diffusion-transformers-NeRFs – Дата звернення: 08.05.2025.

24. TowardsAI. Diffusion Models vs GANs vs VAEs: Comparison of Deep Generative Models. [Електронний ресурс] – Режим доступу до ресурсу: https://pub.towardsai.net/diffusion-models-vs-gans-vs-vaes-comparison-of-deep-generative-models-67ab93e0d9ae – Дата звернення: 08.05.2025.

25. TowardsAI. Diffusion Models vs GANs vs VAEs: Comparison of Deep Generative Models. [Електронний ресурс] – Режим доступу до ресурсу: https://pub.towardsai.net/diffusion-models-vs-gans-vs-vaes-comparison-of-deep-generative-models-67ab93e0d9ae – Дата звернення: 08.05.2025.

26. Exxact Corp. Diffusion and Denoising: Explaining Text-to-Image Generative AI. [Електронний ресурс] – Режим доступу до ресурсу: https://www.exxactcorp.com/blog/deep-learning/diffusion-and-denoising-explaining-text-to-image-generative-ai – Дата звернення: 09.05.2025.

27. Lilian Weng Blog. Diffusion Models. [Електронний ресурс]. – Режим доступу до ресурсу: https://lilianweng.github.io/posts/2021-07-11-diffusion-models/ – Дата звернення: 09.05.2025.

28. MDPI. A Review of Deep Generative Models. [Електронний ресурс] – Режим доступу до ресурсу: https://www.mdpi.com/2224-2708/14/1/17 – Дата звернення: 10.05.2025.

29. ar5iv. A Survey on Generative Models. [Електронний ресурс] – Режим доступу до ресурсу: https://ar5iv.labs.arxiv.org/html/2005.05535 – Дата звернення: 10.05.2025.

30. Alan Zucconi. Understanding the Technology Behind Deepfakes. [Електронний ресурс] – Режим доступу до ресурсу: https://www.alanzucconi.com/2018/03/14/understanding-the-technology-behind-deepfakes/ – Дата звернення: 11.05.2025.

31. Institutional Repository UMSF. [Електронний ресурс] – Режим доступу до ресурсу: http://biblio.umsf.dp.ua/jspui/handle/123456789/4091 – Дата звернення: 11.05.2025.

32. BBC. Technology News. [Електронний ресурс] – Режим доступу до ресурсу: https://www.bbc.com/news/technology-60780142 – Дата звернення: 11.05.2025.

33. Forbes. Бути Віталієм Кличком: як зловмисники створили фейкового мера Києва і провели розмови з керівниками чотирьох європейських столиць. [Електронний ресурс] – Режим доступу до ресурсу: https://forbes.ua/innovations/buti-vitaliem-klichkom-yak-zlovmisniki-stvorili-feykovogo-mera-kieva-i-proveli-rozmovi-z-kerivnikami-chotirokh-evropeyskikh-stolits-29062022-6887 – Дата звернення: 11.05.2025.

34. Суспільне. Deepfake: мери Берліна та Мадрида поспілкувалися з фейковим Кличком. [Електронний ресурс] – Режим доступу до ресурсу: https://suspilne.media/253954-deepfake-meri-berlina-ta-madrida-pospilkuvalisa-z-fejkovim-klickom – Дата звернення: 12.05.2025.

35. NATO Review. Russia's Hybrid War Against the West. [Електронний ресурс] – Режим доступу до ресурсу: https://www.nato.int/docu/review/articles/2024/04/26/russias-hybrid-war-against-the-west – Дата звернення: 12.05.2025.

36. Washington Post. Who spread information/disinformation about the MH17 crash? [Електронний ресурс] – Режим доступу до ресурсу: https://www.washingtonpost.com/news/monkey-cage/wp/2018/09/20/who-spread-information-disinformation-about-the-mh17-crash-we-followed-the-twitter-trail – Дата звернення: 13.05.2025.

37. Euromaidan Press. How Russian troll factory tried to influence Ukraine’s agenda: Analysis of 755,000 tweets. [Електронний ресурс] – Режим доступу до ресурсу: https://euromaidanpress.com/2019/03/14/how-russian-troll-factory-tried-to-influence-ukraines-agenda-analysis-of-755000-tweets – Дата звернення: 13.05.2025.

38. NPR. Russia bot farm: AI disinformation. [Електронний ресурс] – Режим доступу до ресурсу: https://www.npr.org/2024/07/09/g-s1-9010/russia-bot-farm-ai-disinformation – Дата звернення: 13.05.2025.

39. Масовець Ю.І. Використання нейронних мереж для автоматичного розпізнавання тексту. [Електронний ресурс] – Режим доступу до ресурсу: http://dp.knute.edu.ua/jspui/bitstream/123456789/9001/1/%D0%92%D0%9A%D0%A0_%D0%9C%D0%B0%D1%81%D0%BE%D0%B2%D0%B5%D1%86%D1%8C.pdf – Дата звернення: 14.05.2025.

40. Institutional Repository NAU. Artificial Intelligence and Cybersecurity. [Електронний ресурс] – Режим доступу до ресурсу: https://er.nau.edu.ua/items/99a791bd-6174-4127-bc51-44e4e027faf4 – Дата звернення: 15.05.2025.

41. OpenArchive NURE. [Електронний ресурс] – Режим доступу до ресурсу: https://openarchive.nure.ua/entities/publication/7a69ac1b-5eab-4b78-85cd-85892cb44cee – Дата звернення: 16.05.2025.

42. ResearchGate. Evaluation of Federated Learning in Phishing Email Detection. [Електронний ресурс] – Режим доступу до ресурсу:

https://www.researchgate.net/publication/343252836_Evaluation_of_Federated_Learning_in_Phishing_Email_Detection – Дата звернення: 17.05.2025.

43. arXiv. 2202.12233v2. [Електронний ресурс] – Режим доступу до ресурсу: https://arxiv.org/abs/2202.12233 – Дата звернення: 17.05.2025.

44. Домарацький М. Б. Державне управління забезпеченням безпеки критичної інфраструктури в Україні : дис. ... канд. наук з держ. упр.25.00.02. Харків, 2020. – Режим доступу до ресурсу:https://nuczu.edu.ua/images/topmenu/science/spetsializovani-vcheni-rady/disDomarackij.pdf – Дата звернення: 20.05.2025.

45. Жайворонок О. І. Міжнародний досвід протидії інформаційному тероризму та його імплементація в Україні. 2020. [Електронний ресурс] – Режим доступу до ресурсу: http://biblio.umsf.dp.ua/jspui/handle/123456789/4091 – Дата звернення: 20.05.2025.

46. Pappas, S. *Can we trust what we see? AI deepfakes muddy the truth* [Електронний ресурс] // Science News Explores. – 2024. – Режим доступу до ресурсу: https://www.snexplores.org/article/artificial-intelligence-ai-deepfakes-trust-information – Дата звернення: 31 травня 2025 р.

47. Hoxhunt. *AI-Powered Phishing Outperforms Elite Red Teams in 2025* [Електронний ресурс]. – Режим доступу до ресурсу: https://hoxhunt.com/blog/ai-powered-phishing-vs-humans – Дата звернення: 2 червня 2025 р.

48. Van Otten N. How To Implement Anomaly Detection With One-Class SVM In Python : [Електронний ресурс] / Neri Van Otten // Spot Intelligence. – 27 травня 2024 р. – Режим доступу до ресурсу: https://spotintelligence.com/how-to-implement-anomaly-detection-with-one-class-svm-in-python. Дата звернення: 03 червня 2025 р.

ДОДАТКИ

ДОДАТОК А

Табличні дані

Таблиця А-1 Класифікація загроз інформаційній безпеці від ШІ

Категорія загрози	**Тип впливу**	**Потенційні наслідки**	**Рівень складності реалізації**
Генеративні атаки	Створення синтетичного контенту	Дезінформація, фішинг, deepfake	Середній
Adversarial атаки	Маніпуляції з входами моделей	Обхід систем безпеки	Високий
Атаки на приватність	Витік персональних даних	Деанонімізація, профілювання	Середній
Маніпулятивні атаки	Впливи на прийняття рішень	Фінансові шахрайства, політичні маніпуляції	Високий

Таблиця А-2 Порівняння ефективності методів виявлення фішингових повідомлень, згенерованих ШІ

Метод виявлення	**Точність (%)**	**Хибнопозитивні спрацювання (%)**	**Потреба в позначених даних**	**Можливість масштабування**
Класичні сигнатурні фільтри	71.4	13.7	Низька	Обмежена
Логістична регресія + BOW	84.1	9.2	Середня	Висока
BERT-класифікатор (fine-tuned)	94.5	3.8	Висока	Висока
BERT + Federated Learning	93.8	4.1	Низька (локальні дані)	Висока

ChatGPT (zero-shot класифікація)	82.7	11.5	Не потребує	Висока

Продовження додатку А

Таблиця А-3 Приклади ключових ознак фішингових повідомлень, згенерованих ШІ

Тип ознаки	**Опис**	**Приклад**
Синтаксичні шаблони	Повторювані структури речень, притаманні LLM	"На жаль, виникла проблема з вашим обліковим записом..."
Семантична плутанина	Використання надмірної формальності або узагальнення	"З метою покращення нашого сервісу..."
Відсутність персоналізації	Немає згадки про конкретного користувача	"Шановний користувачу..."
Нетипові стилістичні фрагменти	Стиль, схожий на машинний переклад або шаблон	"Будь ласка, виконайте інструкції нижче..."
Вставки коду або нестандартних символів	Некоректне форматування, включення шкідливих скриптів	alert('update');

ДОДАТОК Б

Структурні, процесові та статистичні графічні дані

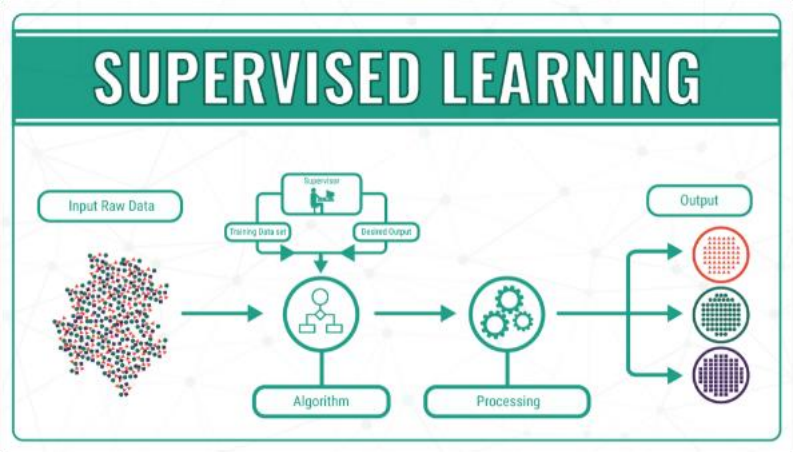

Рисунок Б-1 Графічне відображення супервізованого процесу машинного навчання.

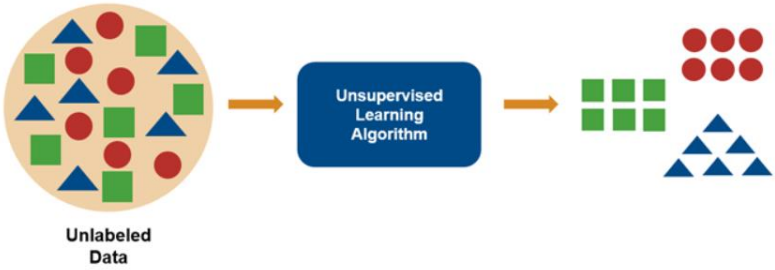

Рисунок Б-2 Графічне відображення процесу несупервізованого навчання

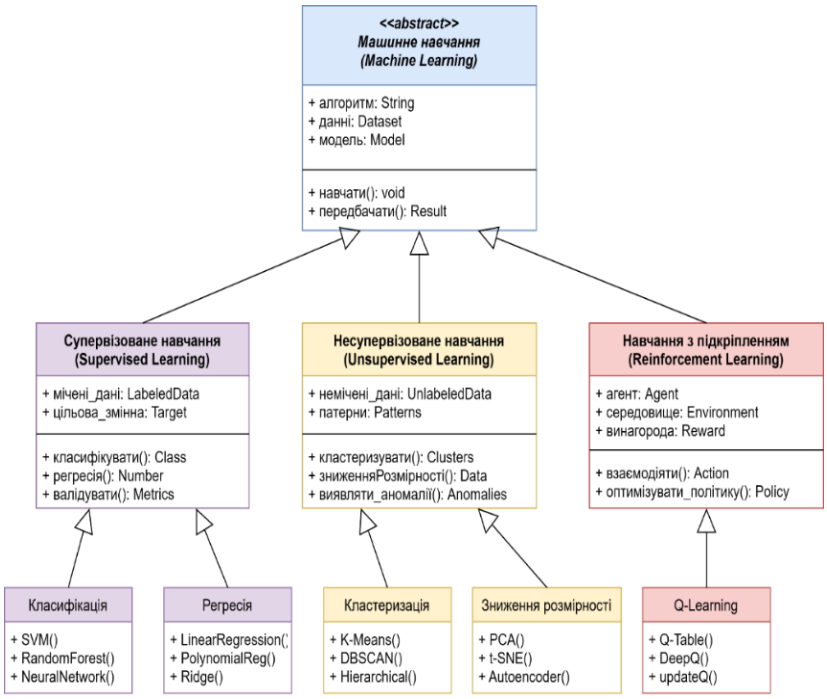

Рисунок Б-3 Графічне відображення архітектури типів машинного навчання

Продовження додатку Б

Рисунок Б-4 Динаміка зростання кількості недостовірних новинних сайтів, згенерованих штучним інтелектом

Рисунок Б-5 Динаміка показників ефективності AI-фішингу порівняно з red team у листопаді 2024 року в розрізі часу навчання користувачів

Продовження додатку Б

Рисунок Б-6 Порівняння ефективності фішингових атак AI-агентів і red team у березні 2025 року залежно від тривалості навчання користувачів

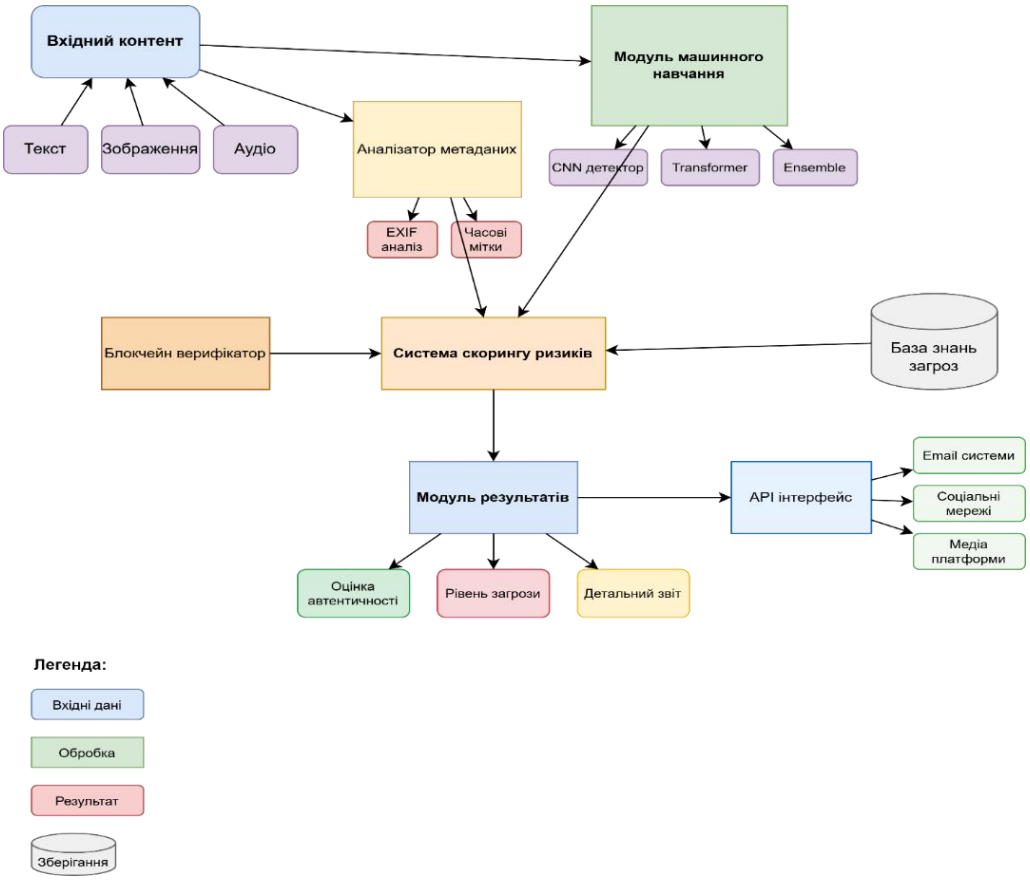

Рисунок Б-7 Архітектура системи виявлення синтетичного контенту

Продовження додатку Б

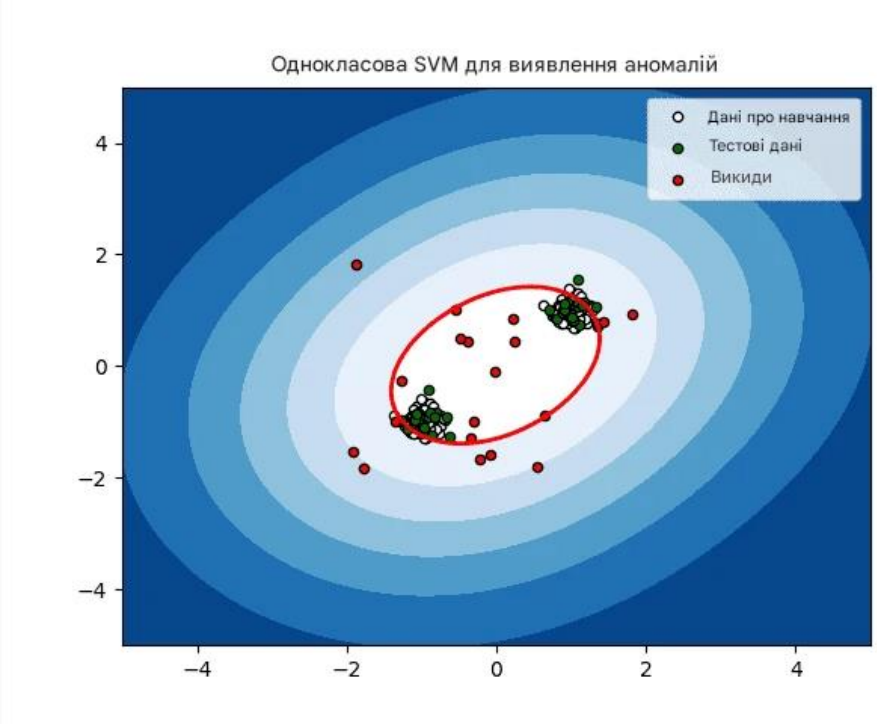

Рисунок Б-8 Візуалізація гіперплощини One-Class SVM та кластеризації нормальних і аномальних точок