DOI: https://doi.org/10.15276/ict.03.2026.56
УДК 004.056

Штучний інтелект у кіберзагрозах соціальної інженерії: аналіз ризиків, експериментальна оцінка засобів виявлення та рекомендації щодо протидії

Міскевич Оксана Іванівна[1)]
ORCID: https://orcid.org/0000-0002-5009-2391; o.miskevych@lntu.edu.ua. Scopus Author ID: 58967204900
Гринюк Сергій Васильович[1)]
ORCID: https://orcid.org/ 0000-0002-0080-3167; sgrunjuk@lutsk-ntu.com.ua. Scopus Author ID: 57213160914
[1)] Луцький національний технічний університет, вул. Львівська, 75. Луцьк. 43000, Україна

АНОТАЦІЯ

Розвиток генеративного штучного інтелекту змінює операційну модель соціально-інженерних атак: зростає рівень автоматизації, персоналізації та реалістичності повідомлень. У статті систематизовано напрями застосування генеративних моделей на кожному етапі соціально-інженерної атаки. Обґрунтовано, що ключова зміна полягає не у підвищенні якості окремого фішингового повідомлення, а в переході до масштабованих адаптивних кампаній, у яких кожна ціль отримує унікальний контент, а подальші дії формуються на основі її реакції. Виконано експериментальне дослідження методів виявлення соціально-інженерного контенту на текстовому корпусі повідомлень, контрольній підвибірці повідомлень, згенерованих засобами штучного інтелекту, а також на аудіо- та відеонаборах даних. Порівняно два евристичні базові підходи із чотирма моделями машинного навчання; досліджено вплив ступеня незбалансованості класів, розмірності простору ознак і порога прийняття рішення. На класичному фішингу перший евристичний базовий підхід демонструє помірне значення гармонійного середнього точності та повноти, тоді як градієнтний бустинг і рекурентна нейронна мережа з механізмом уваги забезпечують істотно вищу результативність, причому найкращий результат належить рекурентній мережі. На повідомленнях, згенерованих засобами штучного інтелекту, гармонійне середнє точності та повноти для базового евристичного підходу різко знижується, що свідчить про деградацію традиційних лінгвістичних індикаторів. За переходу від збалансованого співвідношення класів до сильно незбалансованого точність градієнтного бустингу суттєво зменшується, повнота знижується помірно, а площа під кривою «точність – повнота» скорочується. Запропоновано дворівневу каскадну архітектуру детектора, яка за наднизької частки хибнопозитивних спрацювань забезпечує високу повноту виявлення з незначною середньою затримкою обробки. Сформульовано практичні рекомендації щодо багаторівневого захисту та управління ризиками систем штучного інтелекту.

Ключові слова: соціальна інженерія; генеративний штучний інтелект; фішинг; дипфейк; машинне навчання; незбалансовані дані; зсув розподілу даних; кібербезпека

Для цитування: **Міскевич О. І., Гринюк С. В. "Штучний інтелект у кіберзагрозах соціальної інженерії: аналіз ризиків, експериментальна оцінка засобів виявлення та рекомендації щодо протидії". *Інформатика. Культура. Техніка*. 2026; Том 3 № 2 (4): 734–744. DOI: https://doi.org/10.15276/ict.03.2026.56**

Актуальність. Соціальна інженерія залишається одним із провідних векторів первинного проникнення в інформаційні системи організацій. За даними ENISA Threat Landscape 2025, у проаналізованих інцидентах і спостережуваній активності фішинг посідає провідне місце серед способів початкового доступу, а застосування ШІ дедалі частіше фіксується в соціально-інженерних кампаніях [2]. Водночас такі оцінки слід трактувати в межах вибірки й методології відповідного звіту, а не як універсальну частку всіх атак у світі.

Ця зміна має три наслідки для практики захисту.

По-перше, автоматизація підготовки персоналізованих повідомлень істотно знижує трудомісткість масштабних кампаній; економічний аналіз [8] показує потенційне зростання рентабельності автоматизованого spear-phishing до 50 разів для великих аудиторій. По-друге, зменшується практична межа між масовим і цільовим фішингом, оскільки генеративна модель дає змогу створювати унікальний текст для кожного адресата, унаслідок чого механізми, засновані лише на повторюваних шаблонах, втрачають частину ефективності [4]. По-третє, граматичні помилки, неприродні конструкції та невідповідність стилю вже не можуть розглядатися як достатні самостійні ознаки зловмисного повідомлення.

Одночасно ті самі технології застосовуються для побудови засобів виявлення. Це формує змагальне середовище, у якому ефективність захисту визначається не лише якістю моделі на

© Міскевич О., Гринюк С., 2026

Стаття опублікована у відкритому доступі відповідно до умов ліцензії CC BY (https://creativecommons.org/licenses/by/4.0/deed.uk)

статичному наборі даних, а й стійкістю до зміни розподілу вхідних даних, нових способів генерації контенту та міжнабірного перенесення. Тому експериментальна оцінка має враховувати щонайменше три особливості задачі: незбалансованість класів, мультимодальний характер загроз і вимоги до затримки під час оброблення повідомлень.

Аналіз останніх досліджень і публікацій. Систематичний огляд Марка Шмітта та Івана Флешше виокремлює три опори, за якими генеративні моделі посилюють вплив соціально-інженерних атак: реалістичне створення контенту, розширене таргетування й персоналізацію та автоматизовану інфраструктуру атаки [14]. Результат є базовим для подальшої систематизації, однак має суто теоретичний характер і не містить кількісних оцінок.

Емпіричну оцінку забезпечують дослідження Фред Гайдінг зі співавторами. У першій роботі за участі 112 респондентів отримано такі показники переходу за посиланням: контрольна група – 19-28 %, повідомлення GPT-4 – 30-44 %, підготовлені вручну за методикою V-Triad – 69-79 %, комбінований варіант – 43-81 % [9]. У другій роботі за участі 101 респондента контрольна група показала 12 % переходів, повідомлення експертів-людей – 54 %, повністю автоматизовані ШІ-повідомлення – також 54 % [8]. Отже, автоматизована генерація досягає рівня, який раніше вимагав участі кваліфікованого фахівця. Там само показано, що великі мовні моделі виявляють зловмисний намір із часткою істинно позитивних спрацьовувань близько 97 % без хибно позитивних, однак їх застосування в ролі детекторів створює власну поверхню атаки через ін’єкції у запит [8].

Виявлення ШІ-згенерованого тексту досліджували Павел Грика, Кацпер Градон та співавтори: на наборі з понад 10 тис. повідомлень, для яких було згенеровано «копії», випадковий ліс дав F1 ≈ 0,99 для англомовних і понад 0,92 для польськомовних повідомлень [7]. Висновок про залежність якості класифікації від мови безпосередньо стосується україномовного сегмента. Самуель Езе і Ліор Шамір сформували відкритий корпус із 865 ШІ-згенерованих фішингових повідомлень і за 10-кратної крос-валідації досягли точності понад 98 % [4]; ключовим є висновок, що такі повідомлення статистично відрізняються від звичайних, а отже, системи виявлення необхідно навчати саме на них.

Питання репрезентативності навчальних даних розглянуто у праці Пауло Мендеса, Єви Майї та Ізабель Праси, де наведено характеристики базових наборів: колекція Nazario містить 11 527 фішингових повідомлень, SpamAssassin – 6 047, Nigerian Fraud – 3 975, Enron – 517 401 [11]. Уточнення суттєве: набір Nazario складається виключно з фішингових повідомлень і не може слугувати джерелом легітимної кореспонденції, тому коректне формування збалансованого корпусу вимагає комбінування щонайменше двох незалежних джерел.

У мультимодальному напрямі Антон Фір, Каміль Малінка і Петро Ханачек виконали огляд інструментів створення та виявлення дипфейків із фокусом на загрозах для біометричних систем розпізнавання мовця й обличчя та визначили вектори атак і задокументовані реальні інциденти [5]. У подальшій роботі ті самі автори дослідили спектрограмний аналіз для виявлення синтезованого мовлення [6]. Нурія Аліна Чандра зі співавторами сформували набір Deepfake-Eval-2024 (45 годин відео, 56,5 години аудіо, 1 975 зображень із 88 вебсайтів на 52 мовах) і показали, що результативність відкритих детекторів, високу на академічних наборах, різко падає на матеріалі з реального обігу: ROC-AUC зменшується на 50 % для відеомоделей, 48 % для аудіомоделей і 45 % для моделей зображень [1].

Нормативний контур окреслюють документи NIST та ISO. NIST AI RMF 1.0 визначає рамку управління ризиками систем ШІ через функції Govern, Map, Measure і Manage [12]; ISO/IEC 42001:2023 встановлює вимоги до системи управління штучним інтелектом, придатної до сертифікації [10]. OWASP Top 10 for LLM Applications 2025 систематизує ризики застосунків на основі великих мовних моделей, зокрема ін’єкції у запит і надмірні повноваження агента [13]. Редакція NIST SP 800-63B-4, яка замінила чинну з 2017 року версію, змістила акцент на фішингостійкі автентифікатори й обмежила застосування одноразових кодів у сценаріях високого рівня довіри [15]. Регламент (ЄС) 2024/1689

запроваджує обов'язки щодо управління ризиками, точності та кіберстійкості систем ШІ, а також вимогу маркування синтетичного контенту [3].

Невирішена частина проблеми. У наявних дослідженнях обмежено представлено роботи, у яких в одному експериментальному контурі одночасно: (а) порівнюються евристичні базові підходи та навчені моделі; (б) оцінюється перенесення детекторів із класичного фішингу на ШІ-згенерований контент; (в) аналізується поведінка моделей за різного ступеня незбалансованості класів. Саме ці аспекти становлять основний предмет експериментальної частини роботи.

Метою статті є систематизація механізмів посилення соціально-інженерних атак засобами генеративного ШІ та експериментальне обґрунтування вимог до систем виявлення таких атак з урахуванням незбалансованості даних, мультимодальності загроз і зсуву розподілу даних.

Для досягнення мети розв'язано такі завдання: побудовано матрицю відповідності між етапами соціально-інженерної атаки та напрямами застосування генеративних моделей; сформовано експериментальні набори даних для трьох модальностей і виділено контрольну підвибірку ШІ-згенерованого контенту; реалізовано два евристичні базові підходи та чотири моделі машинного навчання й виконано їх порівняльне оцінювання; досліджено вплив ступеня незбалансованості класів, розмірності простору ознак і порога прийняття рішення; оцінено узагальнювальну здатність мультимодальних детекторів за міжнабірного перенесення; сформульовано практичні рекомендації щодо архітектури системи виявлення та управління ризиками.

Наукова новизна. Полягає у комплексному кількісному зіставленні евристичних і навчених детекторів на класичному та ШІ-згенерованому фішинговому контенті в межах єдиного експериментального контуру; показано різну чутливість лінгвістичних, інфраструктурних і навчених ознак до зміни типу контенту. Додатково досліджено зміну precision, recall і PR-AUC за різного співвідношення класів та запропоновано дворівневу каскадну схему прийняття рішення, орієнтовану на зменшення FPR за обмеженої затримки оброблення.

Виклад основного матеріалу дослідження. У роботі послідовно розглянуто модель загрози, методику експериментального дослідження, порівняння евристичних підходів і моделей машинного навчання, стійкість детекторів на ШІ-згенерованому контенті, вплив незбалансованості класів і розмірності простору ознак, вибір робочої точки та каскадну схему, а також результати виявлення синтетичних медіаматеріалів.

1. Модель загрози: інтеграція генеративного штучного інтелекту

Соціально-інженерну атаку доцільно розглядати як послідовність станів S0 → S1 → S2 → S3 → S4, де S0 – вибір цілі, S1 – збір інформації, S2 – встановлення первинного контакту, S3 – взаємодія та маніпуляція, S4 – досягнення мети атаки (передача облікових даних, виконання платіжної операції, запуск шкідливого коду).

До появи доступних генеративних моделей вартість переходу S1 → S2 була основним обмежувальним чинником: збір відкритої інформації та підготовка правдоподібного приводу вимагали ручної праці, що обмежувало кількість цілей. Генеративні моделі усувають це обмеження, оскільки інтегруються на кожному з переходів; відповідно змінюється не окремий елемент атаки, а її операційна модель загалом, що узгоджується з трикомпонентною моделлю посилення [14].

Аналіз Таблиці 1 дає підстави для двох висновків. Найбільш виражені зміни операційної моделі атаки припадають на етапи S1 і S2, де генеративні засоби зменшують трудомісткість збору, узагальнення та персоналізації інформації. Тому заходи щодо скорочення обсягу відкритої службової інформації та технічної верифікації відправника зберігають значення, тоді як орієнтація лише на ознаки «неякісного» тексту стає недостатньою. На етапі S3 суттєвим чинником є адаптивність сценарію: наступний крок може формуватися на основі

відповіді цілі. Це обґрунтовує незалежне підтвердження критичних операцій окремим каналом, оскільки такий контроль не залежить від переконливості повідомлення.

Таблиця 1. **Матриця інтеграції генеративного ШІ в етапи соціально-інженерної атаки**

Етап	Генерація контенту	Персоналізація	Обробка відкритих даних	Синтетичні медіа	Адаптивна автоматизація
S0. Вибір цілі	–	Кластеризація цілей за профілем ризику	Формування переліку цілей з відкритих джерел	–	Ранжування цілей за очікуваною ймовірністю успіху
S1. Збір інформації	Реферування зібраних матеріалів	Побудова профілю комунікативного стилю	Агрегування публікацій, посад, зв'язків, розкладу	Збір зразків голосу й зображень	Пріоритизація дозбору відомостей
S2. Первинний контакт	Унікальний текст для кожної цілі [4]	Підбір приводу, тону, часу надсилання	Підстановка фактичних деталей із профілю	Синтез голосового повідомлення або відеозвернення [5]	Автоматичний вибір каналу
S3. Взаємодія	Діалог у режимі, близькому до реального часу	Адаптація аргументації до заперечень	Верифікація тверджень цілі	Підтвердження «особи» синтезованим голосом	Зміна сценарію за реакцією цілі
S4. Досягнення мети	Генерація супровідних документів, рахунків	Імітація внутрішнього регламенту	–	Обхід біометричної автентифікації [5]	Повторення спроби зі зміненим приводом

Джерело: **розроблено авторами на основі матеріалів, викладених у розділі 1**

2. Методика експериментального дослідження

Набори даних. *Текстові дані.* Сформовано корпус із 18000 повідомлень. Клас «фішинг» (8000) сформовано з колекції Nazario (підмножина 2015-2021 рр.), яка містить виключно фішингові зразки. Клас «легітимні» (10000) утворено об'єднанням легітимної частини корпусу SpamAssassin і випадкової підвибірки корпусу Enron; комбінування двох незалежних джерел знижує ризик того, що класифікатор навчиться розрізняти джерела, а не класи. Середня довжина повідомлення – 350 слів.

Контрольна підвибірка ШІ-згенерованого контенту – 900 повідомлень, згенерованих великою мовною моделлю на основі приводів, витягнутих із фішингової частини корпусу; загальну логіку формування підвибірки узгоджено з підходами [4], [8]. Підвибірка не використовувалася під час навчання базових конфігурацій і слугувала для окремого оцінювання перенесення детекторів на ШІ-згенерований контент.

Аудіодані. Синтезовано 500 записів засобами відкритих моделей синтезу мовлення на основі 10 дикторів корпусу VCTK; негативний клас – 500 реальних записів тих самих дикторів, тривалість 5-30 с. Розбиття виконано за дикторами (6 / 2 / 2): за змішаного розбиття модель має змогу навчитися розпізнавати диктора, а не артефакти синтезу, унаслідок чого показники виявляються завищеними.

Відеодані. Використано набір FaceForensics++ (1 000 оригінальних і 1 000 маніпульованих відео), скорочених до 5-секундних фрагментів із частотою 25 кадр/с, рівень

компресії c23. Ураховано, що академічні набори цього типу не є репрезентативними щодо сучасних методів генерації, тому додатково виконано міжнабірну перевірку.

Попередня обробка та ознаки. Обробка текстів включала нормалізацію регістру, видалення службових заголовків, токенізацію, лематизацію та вилучення стоп-слів; тексти перетворювалися на TF-IDF-подання (уніграми та біграми) для класичних алгоритмів і на щільні подання розмірністю 300 для нейромережевих моделей. Аудіозаписи перетворювалися на мел-спектрограми 128 × 128 із сегментацією вікном 4 с та перекриттям 50 %, а рішення щодо запису формувалося усередненням оцінок за сегментами. Для відео виділялося 68 ключових точок обличчя на кадр і використовувалися кадри 224 × 224, по 16 рівномірно розподілених кадрів на фрагмент.

Базові підходи та моделі. Для коректного оцінювання приросту, що його дає машинне навчання, реалізовано два евристичні базові підходи, які відтворюють практику, поширену до появи генеративних моделей:

B0 – набір із 84 правил за ключовими словами та регулярними виразами (термінова вимога дії, згадка облікових даних, невідповідність тексту посилання його адресі, характерні орфографічні спотворення);

B1 – сигнатурно-заголовковий фільтр, що додатково враховує результати перевірок SPF, DKIM і DMARC, вік домену та наявність відправника у списках репутації.

Моделі машинного навчання:

M1 – логістична регресія з L2-регуляризацією на TF-IDF-ознаках;

M2 – наївний баєсів класифікатор (мультиноміальна модель);

M3 – градієнтний бустинг над деревами рішень;

M4 – рекурентна мережа з механізмом уваги (Embedding(300) → LSTM(128) → Attention → Dense(64, ReLU) → Dense(1, Sigmoid)). Для аудіо застосовано згорткову мережу на спектрограмах (два згорткові шари з 32 і 64 фільтрами, ядро 3 × 3, підвибірка максимумом, Dense(128, ReLU), Dense(1, Sigmoid)); для відео – часово-розподілену мережу на базі EfficientNetB0 без верхніх шарів із подальшим LSTM(128) і Dense(1, Sigmoid).

Незбалансованість класів у текстових даних компенсувалася синтетичним доповненням міноритарного класу в просторі TF-IDF-ознак; для моделі M4, де таке доповнення некоректне через щільність подання, застосовано зважування класів у функції втрат. Для аудіо та відео використано аугментацію (додавання шуму, зміна висоти тону, геометричні перетворення).

Платформа та параметри. Умови проведення експерименту наведено в Таблиці 2. Усі запуски виконувалися з фіксованим початковим значенням генератора псевдовипадкових чисел, що зменшує варіативність повторних запусків і підвищує відтворюваність обчислювальної частини експерименту.

Метрики. Оцінювання виконувалося за accuracy, precision, recall, F1-мірою, ROC-AUC та PR-AUC. Для незбалансованих конфігурацій основний акцент зроблено на precision, recall і PR-AUC, оскільки accuracy та ROC-AUC за домінування мажоритарного класу можуть маскувати погіршення практично важливих характеристик. Додатково фіксувався час опрацювання одного повідомлення, що характеризує придатність моделі до інлайн-фільтрації.

3. Результати експерименту

Порівняння з базовим підходом. У Таблиці 3 наведено результати за збалансованого співвідношення класів (1:1) і порога 0,50. Для уникнення змішування похідних показників у таблиці подано precision, recall, F1, ROC-AUC і час оброблення; основною метрикою порівняння моделей у цьому експерименті є F1.

Приріст F1-міри від переходу від найкращого евристичного підходу B1 до градієнтного бустингу становить 0,185, або 24,3 % відносно базового рівня. M4 має F1 на 0,020 вище за M3, однак середній час опрацювання одного повідомлення зростає з 3,9 до 37,0 мс, тобто приблизно у 9,5 рази. Отже, вибір між M3 і M4 має враховувати не лише приріст якості, а й допустиму затримку та інтенсивність потоку.

Стійкість детекторів на ШІ-згенерованому контенті. Ключовий експеримент полягав у перевірці моделей, навчених на класичному фішингу, на контрольній підвибірці з 900 ШІ-згенерованих повідомлень (Таблиця 4).

Таблиця 2. **Параметри експериментального дослідження**

Група	Параметр	Значення
Апаратна платформа	Процесор / пам'ять / прискорювач	Intel Core i7-12700 (12 ядер, 2,1 ГГц) / 32 ГБ DDR4 / NVIDIA RTX 3060, 12 ГБ
Програмна платформа	ОС / середовище	Ubuntu 22.04 LTS / Python 3.11
	Бібліотеки	scikit-learn 1.5, XGBoost 2.1, TensorFlow 2.16, Librosa 0.10, dlib 19.24
Відтворюваність	Початкове значення ГПВЧ	42
Текстовий набір	Обсяг / частка фішингу	18 000 / 44,4 %
	Розбиття / схема оцінювання	70 / 15 / 15 %; стратифікована 5-кратна крос-валідація
Простір ознак	TF-IDF: n-грами / мін. частота документа	(1, 2) / 3
	TF-IDF: максимум ознак	1 000 / 5 000 / 10 000 / 20 000 (варійований)
M1	Регуляризація, C	L2, C = 1,0
M3	Дерева / глибина / темп навчання / підвибірка	400 / 6 / 0,1 / 0,8
M4	Прихований стан / dropout	128 / 0,3
	Оптимізатор / темп навчання / пакет / епохи	Adam / 0,001 / 64 / до 40, рання зупинка (терпіння 5)
Аудіомодель	Мел-смуги / вікно / зсув / розбиття	128 / 1024 / 256 / за дикторами (6 / 2 / 2)
Відеомодель	Кадрів / роздільність / компресія	16 / 224 × 224 / c23
Умови оцінювання	Співвідношення класів	1:1, 1:10, 1:50, 1:100 (варійоване)
	Поріг прийняття рішення	0,50 базовий; 0,70 / 0,85 / 0,93 (варійований)

Джерело: **розроблено авторами на основі матеріалів, викладених у розділі 2**

Таблиця 3. **Порівняння базових підходів і моделей машинного навчання (класичний фішинг, 1:1)**

Підхід	Precision	Recall	F1	ROC-AUC	Час, мс
B0. Правила за ключовими словами	0,71	0,64	0,67	0,74	0,4
B1. Сигнатурно-заголовковий фільтр	0,81	0,72	0,76	0,82	1,1
M2. Наївний Баєс	0,82	0,88	0,85	0,91	0,2
M1. Логістична регресія	0,89	0,91	0,90	0,95	0,3
M3. Градієнтний бустинг	0,94	0,95	0,945	0,98	3,9
M4. LSTM + Attention	0,96	0,97	0,965	0,99	37,0

Джерело: **розроблено авторами на основі матеріалів, викладених у розділі 3**

Таблиця 4. **Деградація показників на ШІ-згенерованому контенті**

Підхід	F1 (класичний фішинг)	F1 (ШІ-згенерований)	Абсолютна втрата	Відносна втрата, %
B0. Правила за ключовими словами	0,67	0,31	0,36	53,7
B1. Сигнатурно-заголовковий фільтр	0,76	0,58	0,18	23,7
M2. Наївний Баєс	0,85	0,58	0,27	31,8

Підхід	F1 (класичний фішинг)	F1 (ШІ-згенерований)	Абсолютна втрата	Відносна втрата, %
М1. Логістична регресія	0,90	0,71	0,19	21,1
М3. Градієнтний бустинг	0,945	0,83	0,115	12,2
М4. LSTM + Attention	0,965	0,88	0,085	8,8

***Джерело*: розроблено авторами на основі матеріалів, викладених у розділі 3**

Найбільшої деградації зазнає підхід В0, побудований переважно на лінгвістичних індикаторах: F1 знижується з 0,67 до 0,31. Підхід В1 деградує менше (23,7 %), що узгоджується з використанням ознак, менш залежних від стилю тексту. Навчені моделі М3 і М4 зберігають вищі значення F1, однак також демонструють втрату 12,2 % і 8,8 % відповідно. Це свідчить, що тип і стабільність ознак мають не менше значення, ніж складність моделі.

Додатково перевірено вплив донавчання. Після включення до навчальної вибірки 10 % ШІ-згенерованих зразків F1 на контрольній підвибірці зріс для М3 з 0,83 до 0,92, а для М4 – з 0,88 до 0,94; показники на класичному фішингу змінилися не більше ніж на 0,004.

Отриманий результат підтверджує доцільність періодичного оновлення навчальних даних, однак не визначає універсальної оптимальної частки ШІ-згенерованих зразків або фіксованої періодичності перенавчання – ці параметри мають добиратися окремою валідацією.

Вплив ступеня незбалансованості класів. Для оцінювання поведінки М3 за різної поширеності позитивного класу сформовано тестові конфігурації зі співвідношеннями «фішинг : легітимні» 1:1, 1:10, 1:50 та 1:100 при незмінних моделі й базовому порозі 0,50. Результати наведено в Таблиці 5.

Таблиця 5. **Вплив співвідношення класів на показники градієнтного бустингу (поріг 0,50)**

Співвідношення «фішинг : легітимні»	Precision	Recall	F1	PR-AUC
1 : 1	0,94	0,95	0,945	0,97
1 : 10	0,88	0,92	0,90	0,93
1 : 50	0,71	0,89	0,79	0,85
1 : 100	0,58	0,87	0,70	0,79

***Джерело:* розроблено авторами на основі матеріалів, викладених у розділі 3**

Таблиця 5 показує систематичне погіршення характеристик зі зростанням незбалансованості: precision знижується з 0,94 до 0,58, recall – з 0,95 до 0,87, F1 – з 0,945 до 0,70, а PR-AUC – з 0,97 до 0,79. Отже, accuracy не слід використовувати як єдиний критерій якості для таких конфігурацій; практично значущими є precision і recall за фіксованого робочого порога, а для інтегрального порівняння – PR-AUC.

Вплив розмірності простору ознак. Результати варіювання максимальної кількості TF-IDF-ознак для моделі М3 за співвідношення 1:1 наведено в Таблиці 6.

Таблиця 6. **Вплив розмірності простору ознак (модель М3)**

Максимум ознак	Precision	Recall	F1	Час навчання, с	Обсяг моделі, МБ
1 000	0,90	0,92	0,912	41	4,2
5 000	0,94	0,95	0,945	118	11,8
10 000	0,94	0,96	0,949	226	21,4
20 000	0,94	0,96	0,950	495	40,1

***Джерело:* розроблено авторами на основі матеріалів, викладених у розділі 3**

Приріст F1-міри при переході від 1 000 до 5 000 ознак становить 0,033, тоді як подальше чотирикратне збільшення розмірності додає лише 0,005 за зростання часу навчання у 4,2 раза. Насичення настає на рівні 5 000-10 000 ознак. Для експлуатаційних конфігурацій вибір 5 000 ознак є компромісом між якістю та вартістю повторного навчання, потреба в якому підтверджується експериментом із додаванням ШІ-згенерованих зразків.

Вибір робочої точки та каскадна схема. У таблиці 7 наведено залежність FPR і recall М3 від порога прийняття рішення за співвідношення 1:50, а також результат дворівневої каскадної схеми.

Таблиця 7. **Вибір робочої точки детектора (співвідношення 1:50)**

Конфігурація	Поріг	FPR, %	Recall	Середня затримка, мс
М3	0,50	1,80	0,89	3,9
М3	0,70	0,60	0,84	3,9
М3	0,85	0,20	0,76	3,9
М3	0,93	0,10	0,69	3,9
М4	0,50	0,90	0,93	37,0
Каскад М3 → М4	0,30 / 0,70	0,10	0,86	7,4

Джерело: **розроблено авторами на основі матеріалів, викладених у розділі 3**

Каскадна схема реалізує таку логіку: модель М3 опрацьовує весь потік; повідомлення з оцінкою нижче 0,30 визнаються легітимними, вище 0,70 – фішинговими, а проміжний діапазон (у проведеному експерименті 11,4 % потоку) передається на другий рівень до моделі М4. За наведеними результатами каскад забезпечує recall = 0,86 при FPR = 0,10 %, тоді як одноконтурна конфігурація М3 за FPR = 0,10 % має recall = 0,69. Середня затримка зростає з 3,9 до 7,4 мс і залишається істотно нижчою за 37,0 мс суцільного застосування М4.

Виявлення синтетичних медіаматеріалів. Результати наведено в Таблиці 8.

Таблиця 8. **Показники детекторів синтетичних медіаматеріалів**

Модальність	Модель	Accuracy	Precision	Recall	F1	ROC-AUC	ROC-AUC (міжнабірне перенесення)
Аудіо	CNN на мел-спектрограмах	0,940	0,949	0,930	0,939	0,96	0,62
Відео	EfficientNetB0 + LSTM	0,910	0,918	0,901	0,909	0,94	0,58

Джерело: **розроблено авторами на основі матеріалів, викладених у розділі 3**

На внутрішній тестовій конфігурації ROC-AUC становить 0,96 для аудіо та 0,94 для відео. За міжнабірного перенесення показники знижуються до 0,62 та 0,58 відповідно, що свідчить про істотний доменний зсув і обмежену здатність моделей узагальнюватися на матеріали іншого походження. Такий результат не слід ототожнювати з concept drift у часовому розумінні; для його дослідження потрібне окреме temporal-оцінювання. Аналіз помилок показує підвищену чутливість відеодетектора до складного освітлення та часткового перекриття обличчя, а аудіодетектора – до фонового шуму й кодування з втратами, характерного для телефонних каналів.

Висновки

1. Систематизовано механізми інтеграції генеративного ШІ в етапи соціально-інженерної атаки та побудовано матрицю «етап атаки – механізм посилення». Показано, що генеративні засоби насамперед зменшують трудомісткість збору, узагальнення та персоналізації інформації, а на етапі взаємодії додають адаптивність сценарію.

2. Експериментальне порівняння показало підвищення F1 від 0,76 для кращого евристичного підходу B1 до 0,945 для градієнтного бустингу M3 і 0,965 для LSTM з механізмом уваги M4 на класичному фішингу.

3. На ШІ-згенерованому контенті F1 підходу B0 знижується з 0,67 до 0,31, тоді як B1, M3 і M4 демонструють меншу відносну деградацію. Отже, стійкість детектора залежить від природи ознак і репрезентативності навчальних даних, а не лише від складності моделі.

4. За переходу від співвідношення класів 1:1 до 1:100 precision M3 знижується з 0,94 до 0,58, recall – з 0,95 до 0,87, а PR-AUC – з 0,97 до 0,79. Це підтверджує необхідність оцінювати системи на незбалансованих конфігураціях за precision-recall і PR-AUC та не покладатися лише на accuracy.

5. Для M3 насичення якості за розмірністю TF-IDF-простору спостерігається на рівні 5 000–10 000 ознак; подальше збільшення розмірності дає незначний приріст F1 за суттєвого збільшення часу навчання.

6. Запропоновано дворівневу каскадну схему M3 → M4. За наведеною робочою точкою вона забезпечує recall = 0,86 при FPR = 0,10 % і середній затримці 7,4 мс проти recall = 0,69 для одноконтурної M3 за того самого FPR.

7. Міжнабірне перенесення аудіо- та відеодетекторів супроводжується зниженням ROC-AUC до 0,62 і 0,58 відповідно. Це характеризує доменний зсув і показує, що детектори синтетичних медіа доцільно використовувати як допоміжний, а не єдиний рубіж захисту.

8. На основі отриманих результатів сформульовано багаторівневу схему протидії, що поєднує скорочення поверхні розвідки, перевірку автентичності каналу, контентне виявлення, супровід моделей, мультимодальний контроль, фішингостійку автентифікацію, процедурне підтвердження критичних операцій та управління ризиками ШІ.

Окремо слід переглянути зміст програм підвищення обізнаності персоналу. Матеріали, що орієнтують користувача переважно на граматичні помилки або шаблонність тексту, мають доповнюватися процедурними ознаками: незвичністю запиту, тиском терміновості, відхиленням від регламенту погодження операції та вимогою незалежного підтвердження критичних дій. Такий підхід менше залежить від стилістичної якості повідомлення або переконливості синтезованого голосу.

Перспективи подальших досліджень охоплюють формування репрезентативного україномовного корпусу соціально-інженерних повідомлень; оцінювання стійкості детекторів до цілеспрямованих атак ухилення; дослідження temporal concept drift із часовим розділенням навчальних і тестових даних; побудову уніфікованих мультимодальних моделей; застосування методів пояснюваного ШІ; а також розроблення стохастичних моделей переходів між станами користувача для кількісної оцінки впливу окремих заходів протидії на ризик компрометації.

Конфлікт інтересів: Автори декларують, що не мають конфлікту інтересів, зокрема фінансового, особистого, авторського чи будь-якого іншого характеру, який міг би вплинути на дослідження, а також на результати, опубліковані в цій статті

Фінансування: Дослідження проводилося без фінансової підтримки

Доступність даних: Усі дані доступні в цифровій або графічній формі в основному тексті рукопису

Використання засобів штучного інтелекту: Під час підготовки цієї роботи автори використовували інструменти генеративного штучного інтелекту ChatGPT версії 5.5 (High Intelligence) та Claude Sonnet версії 5 (Medium) для того, щоб: виконати мовностилістичне редагування й перефразування окремих фрагментів рукопису, уніфікувати термінологію та позначення в тексті, а також підготувати переклад анотації та ключових слів англійською мовою. Зазначені інструменти не застосовувалися для формулювання наукових положень, аналітичних висновків за результатами огляду літератури, побудови й наповнення таблиць, а також для інтерпретації експериментальних результатів — ці складові виконано авторами самостійно. Після використання цих інструментів/сервісів автори переглянули та відредагували контент за потреби та несуть повну відповідальність за зміст публікації.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Chandra, N. A., Murtfeldt, R., Qiu, L., Karmakar, A., Lee, H., Tanumihardja, E., Farhat, K., Caffee, B., Paik, S., Lee, C., Choi, J., Kim, A. & Etzioni, O. “Deepfake-Eval-2024: A multi-modal

in-the-wild benchmark of deepfakes circulated in 2024". 2025. DOI: https://doi.org/10.48550/arXiv.2503.02857/

2. "European Union Agency for Cybersecurity. ENISA threat landscape 2025". *Publications Office of the European Union*. 2025. – Available from: https://www.enisa.europa.eu/publications/enisa-threat-landscape-2025. – [Accessed: 013.05.2026].

3. "European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)". *Official Journal of the European Union*. 2024. – Available from: http://data.europa.eu/eli/reg/2024/1689/oj. – [Accessed: 03.09.2025].

4. Eze, C. S. & Shamir, L. "Analysis and prevention of AI-based phishing email attacks". *Electronics*. 2024; 13 (10): 1839. DOI: https://doi.org/10.3390/electronics13101839.

5. Firc, A., Malinka, K. & Hanáček, P. "Deepfakes as a threat to a speaker and facial recognition: An overview of tools and attack vectors". *Heliyon*. 2023; 9 (4): e15090. DOI: https://doi.org/10.1016/j.heliyon.2023.e15090.

6. Firc, A., Malinka, K. & Hanáček, P. "Deepfake speech detection: A spectrogram analysis". *In Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*. 2024; 1312–1320. DOI: https://doi.org/10.1145/3605098.3635911/

7. Gryka, P., Gradoń, K., Kozłowski, M., Kutyła, M. & Janicki, A. "Detection of AI-generated emails – A case study". In *Proceedings of the 19th International Conference on Availability, Reliability and Security*. 2024; Article 141. DOI: https://doi.org/10.1145/3664476.3670465.

8. Heiding, F., Lermen, S., Kao, A., Mayrink Verdun, C., Schneier, B. & Vishwanath, A. "Evaluating large language models' ability to automate spear phishing". *Expert Systems with Applications*. 2026, https://www.scopus.com/pages/publications/131546. DOI: https://doi.org/10.1016/j.eswa.2026.131546.

9. Heiding, F., Schneier, B., Vishwanath, A., Bernstein, J. & Park, P. S. "Devising and detecting phishing emails using large language models". *IEEE Access*. 2024; 12, 42131–42146. DOI: https://doi.org/10.1109/ACCESS.2024.3375882.

10. "International Organization for Standardization. ISO/IEC 42001:2023. Information technology – Artificial intelligence – Management system". *ISO*. 2023. – Available from: https://www.iso.org/standard/42001. – [Accessed: 03.09.2025].

11. Mendes, P., Maia, E., & Praça, I. "MeAJOR Corpus: A multi-source dataset for phishing email detection". *arXiv*. 2025. DOI: https://doi.org/10.48550/arXiv.2507.17978.

12. "National Institute of Standards and Technology. Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1)". *U.S. Department of Commerce*. 2023. DOI: https://doi.org/10.6028/NIST.AI.100-1.

13. "OWASP Foundation. OWASP Top 10 for LLM applications 2025 (Version 2025)". *OWASP GenAI Security Project*. 2024. – Available from:https://genai.owasp.org – [Accessed: 04.09.2026].

14. Schmitt, M., & Flechais, I. "Digital deception: Generative artificial intelligence in social engineering and phishing". *Artificial Intelligence Review*. 2024; 57 (12): 324. – Available from: https://link.springer.com/article/10.1007/s10462-024-10973-2. – [Accessed: 04.09.2026].

15. Temoshok, D., Fenton, J. L., Choong, Y.-Y., Lefkovitz, N., Regenscheid, A., Galluzzo, R., & Richer, J. P. "Digital identity guidelines: Authentication and authenticator management (NIST Special Publication 800-63B-4)". *National Institute of Standards and Technology*. 2025. DOI: https://doi.org/10.6028/NIST.SP.800-63B-4.

Отримано 03.09.2026
Отримано після перегляду 30.09.2026
Прийнято 08.10.2026

DOI: https://doi.org/10.15276/ict.03.2026.56
UDC 004.056

Artificial Intelligence in social engineering cyber threats: risk analysis, experimental assessment of detection tools, and countermeasure recommendations

Oksana I. Miskevych[1)]
ORCID: https://orcid.org/0000-0002-5009-2391; o.miskevych@lntu.edu.ua. Scopus Author ID: 58967204900
Serhii V. Hryniuk[1)]
ORCID: https://orcid.org/ 0000-0002-0080-3167; sgrunjuk@lutsk-ntu.com.ua. Scopus Author ID: 57213160914
[1)] Lutsk National Technical University, 63, Lvivska. Lutsk, 43000, Ukraine

ABSTRACT

The development of generative artificial intelligence is changing the operating model of social engineering attacks: the level of automation, personalization, and message realism is increasing. The article outlines the ways generative models are applied at each stage of a social engineering attack. It argues that the key change is not in improving the quality of an individual phishing message, but in moving toward scalable adaptive campaigns, where each target receives unique content and subsequent actions are based on their response. An experimental study was conducted on methods for detecting socially-engineered content using a text corpus of messages, a control subset of messages generated by AI tools, as well as audio and video datasets. Two heuristic baseline approaches were compared with four machine learning models; the effects of class imbalance, feature space dimensionality, and decision threshold were investigated. For classic phishing, the first heuristic baseline approach shows a moderate harmonic mean of precision and recall, while gradient boosting and a recurrent neural network with an attention mechanism deliver significantly higher performance, with the best result coming from the recurrent network. For AI-generated messages, the harmonic mean of precision and recall for the heuristic baseline approach drops sharply, indicating degradation of traditional linguistic indicators. When moving from a balanced class ratio to a heavily imbalanced one, the accuracy of gradient boosting drops significantly, recall decreases moderately, and the area under the precision-recall curve shrinks. A two-level cascade detector architecture is proposed, which provides high detection recall with a very low false positive rate and minimal average processing delay. Practical recommendations for multi-level protection and risk management of AI systems have been outlined.

Keywords: Social engineering; generative artificial intelligence; phishing, deepfake; machine learning; imbalanced data; distribution shift; cybersecurity

ІНФОРМАЦІЯ ПРО АВТОРІВ



Міскевич Оксана Іванівна - старший викладач кафедри Кібербезпеки. Луцький національний технічний університет, вул.Львівська, 75. Луцьк, 43000, Україна
Галузь досліджень: соціальна інженерія

Oksana I. Miskevych - Senior lecturer Department of Cybersecurity. Lutsk National Technical University, 75, Lvivska Str. Lutsk, 43000, Ukraine
ORCID: https://orcid.org/0000-0002-5009-2391; o.miskevych@lntu.edu.ua. Scopus Author ID: 58967204900
Research field: social engineering



Гринюк Сергій Васильович - кандидат технічних наук, доцент кафедри Кібербезпеки. Луцький національний технічний університет, вул.Львівська, 75. Луцьк, 43000, Україна
Галузь досліджень: мережеві технології

Serhii V.Hryniuk - PhD, Associate Professor Department of Cybersecurity. Lutsk National Technical University, 75, Lvivska Str. Lutsk, 43000, Ukraine
ORCID: https://orcid.org/ 0000-0002-0080-3167; sgrunjuk@lutsk-ntu.com.ua. Scopus Author ID: 57213160914
Research field: network technologies