Вплив перефразування повідомлень цільового фішингу великими мовними моделями на спрацювання правил поштових фільтрів

Impact of LLM Rephrasing of Spear-Phishing Emails on Email Filter Rule Activation

Роман Жишко
Corresponding author: студент кафедри "Захист інформації", e-mail: roman.zhyshko.mkbas.2025@lpnu.ua, ORCID ID: https://orcid.org/0009-0001-9188-3931

Єлизавета Черемних
студент кафедри "Захист інформації", e-mail: yelyzaveta.cheremnykh.mkbas.2025@lpnu.ua, ORCID ID: https://orcid.org/0009-0005-8901-7101

Віталій Сусукайло
доктор філософії, старший викладач кафедри "Захист інформації", e-mail: vitalii.a.susukailo@lpnu.ua, ORCID ID: https://orcid.org/0000-0003-4431-9964

Національний університет "Львівська політехніка", Львів, Україна

Roman Zhyshko
Corresponding author: Student, Department of Information Security, e-mail: roman.zhyshko.mkbas.2025@lpnu.ua, ORCID ID: https://orcid.org/0009-0001-9188-3931

Yelyzaveta Cheremnykh
Student, Department of Information Security, e-mail: yelyzaveta.cheremnykh.mkbas.2025@lpnu.ua, ORCID ID: https://orcid.org/0009-0005-8901-7101

Vitalii Susukailo
PhD, Senior Lecturer, Department of Information Security, e-mail: vitalii.a.susukailo@lpnu.ua, ORCID ID: https://orcid.org/0000-0003-4431-9964

Lviv Polytechnic National University, Lviv, Ukraine

Received: August 28, 2026 | Revised: September 30, 2026 | Accepted: October 31, 2026

УДК 004.8:004.056.5

DOI: https://doi.org/10.33445/sds.2026.16.5.16

Мета дослідження. Визначити вплив перефразування повідомлень цільового фішингу великими мовними моделями на спрацьовування правил поштових фільтрів у лабораторних умовах.

Методи дослідження. Проведено експеримент із трьома групами по 150 повідомлень: фішинговими листами, їхніми перефразованими версіями та легітимними листами. Використано набори даних ePhishLLM, Nazario та Enron. Після видалення дублікатів мережеві ідентифікатори замінено стандартизованими заповнювачами. Локальна мовна модель перефразувала листи, зберігши зміст і запитувані дії. Уніфіковані повідомлення перевірено фільтрами SpamAssassin і Rspamd; проаналізовано оцінки та спрацювання правил.

Результати дослідження. Rspamd класифікував усі контрольні листи як підозрілі через особливості лабораторних заголовків, що унеможливило порівняння груп за двійковою класифікацією. SpamAssassin виявив два початкові й один перефразований фішинговий лист. Після перефразування кількість спрацьовувань правил, пов'язаних із терміновістю, імітацією брендів і фішинговими адресами, зменшилася відповідно з 6 до 0, з 3 до 0 та з 2 до 1. Частота спрацьовування правил щодо облікових записів і безпекових дій не зменшилася.

Теоретична цінність дослідження. Обґрунтовано інформативність аналізу окремих правил порівняно з двійковою класифікацією.

Практична цінність дослідження. Підхід підходить для оцінювання стійкості контентних евристик до стилістичних змін.

Оригінальність / цінність дослідження. Виявлено зміни в профілях спрацьовування правил для пар початкових і перефразованих повідомлень.

Обмеження / напрями подальших досліджень. Результати обмежені лабораторними умовами; необхідна перевірка моделей і конфігурацій фільтрів.

Тип статті: емпірична.

Purpose. To assess how large language model rephrasing affects email-filter rule activation for spear-phishing messages under controlled laboratory conditions.

Method. A controlled experiment examined three groups of 150 messages: original phishing emails, rephrased versions, and legitimate business correspondence. Data were obtained from ePhishLLM, Nazario, and Enron. Following deduplication, network identifiers were replaced with standardized placeholders. A local language model rephrased messages while preserving their intent and requested actions. Standardized emails were evaluated using SpamAssassin and Rspamd, with a focus on classification outcomes and rule activation.

Results. Rspamd incorrectly flagged all legitimate control messages because of laboratory header characteristics, preventing meaningful binary comparisons. SpamAssassin detected two original and one rephrased phishing email. Following rephrasing, activations associated with urgency, brand impersonation, and phishing-related addresses decreased from 6 to 0, 3 to 0, and 2 to 1, respectively. Account- and security-related rules were maintained or had their activation frequency increased.

Theoretical implications. Rule-level analysis reveals changes obscured by binary classification metrics.

Practical implications. The approach supports laboratory evaluation of the robustness of content-based filtering.

Originality / value. The study identifies measurable changes in rule activation across paired phishing messages.

Limitations / future research. Findings are laboratory-specific and require validation across additional models and configurations.

Article type. Empirical research.

Ключові слова: аналіз правил; контентна фільтрація; корпоративна електронна пошта; стилістична нормалізація; ознаки соціальної інженерії; автоматизоване виявлення загроз.

Key words: filter-rule analysis; content-based filtering; corporate email; stylistic normalization; social-engineering indicators; automated threat detection.

Вступ

Фішингові атаки залишаються одним із найпоширеніших каналів соціальної інженерії. Їхня ефективність пояснюється не лише технічним механізмом доставки листа, а й тим, що повідомлення змушує користувача виконати дію: перейти за посиланням, відкрити файл, підтвердити дані, здійснити платіж або надати облікові дані. Особливо небезпечними є повідомлення типу spear phishing (цільового фішингу), які імітують звичайну ділову комунікацію, використовують контекст організації, службові ролі, фінансові або документні процеси. Масштаб проблеми підтверджується регулярними звітами APWG, у яких фішинг розглядається як стійка та масова форма кібершахрайства [1].

Поява великих мовних моделей змінює характер цієї загрози. Традиційні фішингові листи часто містили ознаки, які легко помітити: граматичні помилки, неприродний стиль, надмірну терміновість, грубі спроби імітації бренду або незграбні формулювання. LLM дають змогу швидко перефразувати такий текст у більш природний і формально-корпоративний стиль. У результаті частина поверхневих маркерів, на які можуть орієнтуватися фільтри на основі правил або евристик, послаблюється, хоча сам ризик-релевантний запит може залишатися незмінним.

SpamAssassin і Rspamd, як і багато інших класичних поштових фільтрів, оцінюють лист за сукупністю ознак. Вони можуть враховувати заголовки, тіло повідомлення, текстові евристики, бальну оцінку за правилами, байєсівську або статистичну фільтрацію, DNSBL/URIBL, автентифікаційні та репутаційні перевірки. Такий підхід корисний у реальній поштовій інфраструктурі, але він ускладнює ізольоване дослідження саме текстового впливу LLM-перефразування. Якщо рішення фільтра значною мірою залежить від SPF/DKIM/DMARC, IP-репутації, URL-репутації або SMTP-контексту, то підсумкова оцінка може відображати технічну якість доставки, а не зміст листа.

Тому в цій роботі увага зосереджена не лише на факті виявлення повідомлення, а й на тому, які саме правила або символи спрацьовують до та після перефразування. Такий аналіз дає змогу перевірити, чи справді після LLM-перефразування зникають явні сигнатурні маркери фішингу: термінова бізнес-лексика, бренд-фішингові патерни, підозрілі URI-ознаки. Окремо важливо з’ясувати, чи зберігається при цьому семантика, пов’язана з обліковими записами, документами або безпекою повідомлення.

Метою статті є оцінювання впливу LLM-перефразування повідомлень типу spear phishing на спрацьовування поштових фільтрів SpamAssassin і Rspamd в очищеному середовищі, зосередженому на контенті.

Для досягнення мети поставлено такі завдання: сформувати набір оригінальних повідомлень із потенційними ознаками фішингу P0 і контрольну групу легітимних ділових листів H0; згенерувати парний набір P1 шляхом контрольованого LLM-перефразування P0; перевірити збереження ризик-релевантної семантики у P1; підготувати уніфіковані EML-повідомлення без зовнішніх інфраструктурних чинників; провести оцінювання SpamAssassin і Rspamd; проаналізувати не лише двійкову метрику виявлення, а й зміни профілю спрацьовування правил і символів фільтрів.

Наукова новизна дослідження полягає в експериментальному встановленні змін у профілі спрацьовування окремих правил SpamAssassin до та після LLM-перефразування парних повідомлень P0–P1. Зафіксовано зменшення кількості спрацювань URG_BIZ (з 6 до 0), FUZZY_PAYPAL (з 3 до 0) і URI_PHISH (з 2 до 1), тоді як TVD_PH_BODY_ACCOUNTS_PRE (з 2 до 4), TVD_PH_SEC (з 2 до 3) і TVD_PH_SUBJ_META1 (з 7 до 9) спрацьовували частіше. Показано, що аналіз профілю окремих правил дає змогу виявляти зміни текстових ознак, які не відображаються безпосередньо у двійковій метриці виявлення.

Огляд літератури

Проблематика фішингу та соціальної інженерії широко розглядається у звітах галузевих і міжнародних організацій. Звіти APWG демонструють сталу активність фішингових кампаній і зростання складності шахрайських сценаріїв, зокрема компрометації ділового листування (business email compromise) [1]. У звітах ENISA фішинг і соціальна інженерія також віднесені до важливих складових сучасного ландшафту кіберзагроз, що впливають як на приватний, так і на корпоративний сектор [8].

Окремий напрям досліджень стосується ролі генеративного штучного інтелекту в соціальній інженерії. У працях, присвячених цифровому обману (digital deception) та соціально-інженерним атакам із використанням ШІ (AI-enabled social engineering), підкреслюється, що генеративні моделі можуть підвищувати реалістичність атак, масштабувати створення переконливих повідомлень і спрощувати персоналізацію контенту [26; 9]. Водночас значна частина робіт зосереджується або на генерації фішингових повідомлень, або на застосуванні машинного навчання та LLM для їх виявлення [30; 15; 28]. Менш розкритим залишається питання того, як саме LLM-перефразування впливає на реакцію класичних поштових фільтрів на рівні окремих правил і сигнатур.

Для такого аналізу практичний інтерес становлять SpamAssassin і Rspamd. Обидва інструменти надають не лише підсумкову оцінку або рішення, а й деталізовану інформацію про правила чи символи, які спрацювали для кожного повідомлення [2; 25]. Це дає змогу розглядати не лише результат класифікації, а й механізм, за якого фільтр дійшов відповідного рішення.

Сучасні дослідження розглядають як переконливість повідомлень, згенерованих великими мовними моделями, так і можливості їх виявлення. Heiding et al. [13] порівняли згенеровані листи з повідомленнями, створеними за експертною методикою, та показали, що результати залежать від сценарію й способу персоналізації. Fairbanks and Serra [10] дослідили перефразування фішингових листів із збереженням семантики та індикаторів компрометації, а Han et al. [12] використали оцінки SpamAssassin для послідовного вдосконалення згенерованих повідомлень. Opara et al. [22] безпосередньо порівняли фільтри Gmail, Outlook і Yahoo та стилометричні методи для виявлення фішингових листів, згенерованих штучним інтелектом. Їхня робота є близькою за тематикою до цього дослідження, однак у ній оцінювалися загальні результати роботи поштових сервісів, тоді як тут аналізується профіль спрацьовування конкретних правил у контрольованому середовищі з мінімізованим впливом інфраструктурних чинників.

Паралельно розвиваються підходи до виявлення фішингу за допомогою великих мовних моделей. Koide et al. [16] запропонували засіб для класифікації фішингових вебсторінок, а Koide et al. [17] дослідили застосування великих мовних моделей для виявлення фішингових електронних листів. Aseeri and Bohacek [3] оцінили ансамблевий підхід із використанням кількох мовних моделей. Gutierrez et al. [11] розглянули здатність стилометричних детекторів узагальнювати ознаки між різними генеративними моделями. Ці результати підкреслюють потребу перевіряти стійкість підходів у різних умовах і доповнювати загальну класифікацію аналізом окремих ознак повідомлень.

Методологія дослідження

Цільовий фішинг (spear phishing) – це різновид фішингової атаки, у якій повідомлення орієнтується не на випадкового отримувача, а на конкретну особу, групу працівників або організацію. Такі листи зазвичай імітують звичну ділову взаємодію: обговорення документа, платіж, кадрове повідомлення, запит від партнера, перевірку облікового запису або доступ до внутрішнього ресурсу. Їхня небезпека полягає не лише в технічному каналі доставки, а й у

використанні прийомів соціальної інженерії (social engineering), коли користувача спонукають виконати дію, вигідну зловмиснику [19; 18].

Для цільового фішингу важливою є не окрема "підозріла" фраза, а поєднання кількох ознак: наявність службового контексту, конкретного адресата або ролі, згадка про бізнес-процес, документ, платіж, обліковий запис чи іншу дію, що потребує реакції з боку отримувача. Класичний масовий фішинг часто видає себе грубими мовними помилками, надмірною терміновістю, неприродним стилем або очевидною імітацією відомого бренду [27; 5]. У цільових листах ці ознаки можуть бути значно менш вираженими. Саме тому просте виявлення окремих слів-маркерів не завжди достатньо для оцінювання таких повідомлень.

Великі мовні моделі (large language models, LLM) додатково ускладнюють цю проблему. Вони здатні швидко виправляти граматику, згладжувати надмірно емоційний тон, змінювати структуру речень і перетворювати текст на більш нейтральний корпоративний стиль. Для соціально-інженерних атак це означає, що повідомлення може втратити частину поверхневих маркерів фішингу, але зберегти основний запит: перейти на ресурс, надати дані, переглянути документ або підтвердити дію. У сучасних дослідженнях генеративний штучний інтелект розглядається як чинник, що може підвищувати реалістичність і масштабованість соціально-інженерних сценаріїв [26; 13; 24].

У цій роботі LLM-перефразування не розглядається як генерація нової атаки з нуля. Йдеться про контрольовану стилістичну трансформацію вже наявного фішингового кандидата. Для такого перефразування принциповими є дві умови: по-перше, збереження семантичного наміру, запитуваної дії та ризик-релевантних ознак; по-друге, нормалізація явних маркерів психологічного тиску, терміновості, граматичних помилок і неприродного стилю. Така постановка дає змогу перевірити, чи реагують поштові фільтри на зміст фішингового запиту, чи переважно на його поверхневе оформлення.

Поштові фільтри SpamAssassin і Rspamd належать до класу систем, які оцінюють повідомлення за сукупністю багатьох ознак. Сучасні підходи до автоматизованого виявлення фішингових повідомлень також використовують поєднання текстових, структурних та інших ознак [4; 6]. SpamAssassin використовує бальну модель оцінювання (scoring framework): до повідомлення застосовується набір правил і тестів, кожен із яких може додавати або віднімати певну кількість балів. До таких перевірок належать аналіз заголовків, тіла листа, статистичні й евристичні тести, байєсівська фільтрація, перевірки за списками блокування доменів або адрес та інші механізми [2]. Підсумкова оцінка визначає, чи буде лист класифіковано як небажаний або підозрілий.

Rspamd працює за схожою логікою, але оперує системою символів. Кожен символ відповідає певній ознаці повідомлення: технічній, репутаційній, контентній або пов'язаній із політикою обробки. Підсумковий бал порівнюється з порогами, які визначають рекомендовану дію: пропустити лист, додати заголовок, застосувати відкладену доставку або відхилити повідомлення [25]. Через це результат роботи Rspamd залежить не лише від тексту листа, а й від технічного контексту: автентифікації, SMTP-ланцюга, доменів, заголовків, MIME-структури та репутаційних перевірок.

Для оцінювання впливу LLM-перефразування ці позатекстові чинники потрібно контролювати. Якщо залишити реальні URL-адреси, домени, IP-репутацію, SPF/DKIM/DMARC або повний SMTP-контекст, зміна результатів фільтра може бути пов'язана не з перефразуванням, а з інфраструктурними показниками доставки. Тому в дослідженні використовується очищене середовище, зосереджене на контенті (sanitized/content-only environment): URL-адреси та email-адреси замінюються стабільними токенами-заповнювачами, а EML-повідомлення формуються з уніфікованими технічними заголовками. Такий режим не претендує на повну імітацію реальної поштової інфраструктури, але дає змогу окремо дослідити текстовий компонент.

Оцінювання лише за двійковим показником виявлення (binary detection rate) у такій постановці може бути недостатнім. Якщо фільтр майже не реагує на текстові ознаки або, навпаки, спрацьовує через технічні характеристики всіх повідомлень, лише результат класифікації “виявлено/не виявлено” мало пояснює механізм ухвалення рішення. Тому в роботі додатково використовується аналіз на рівні правил і символів. Він показує, які саме ознаки активувалися до та після перефразування, і дає змогу виявити зміни, що не завжди помітні за загальною метрикою виявлення.

Методика дослідження побудована як контрольований експеримент із трьома групами електронних повідомлень: P0, P1 та H0. Група P0 містить оригінальні початкові фішингові кандидати з ознаками цільового фішингу, група P1 – їхні перефразовані версії, а H0 – контрольну групу легітимних ділових листів. Порівняння P0 і P1 дає змогу оцінити, як змінюється реакція поштових фільтрів після LLM-перефразування, а група H0 використовується для контролю хибних спрацьовувань. Загальну послідовність етапів дослідження наведено на рис. 1.

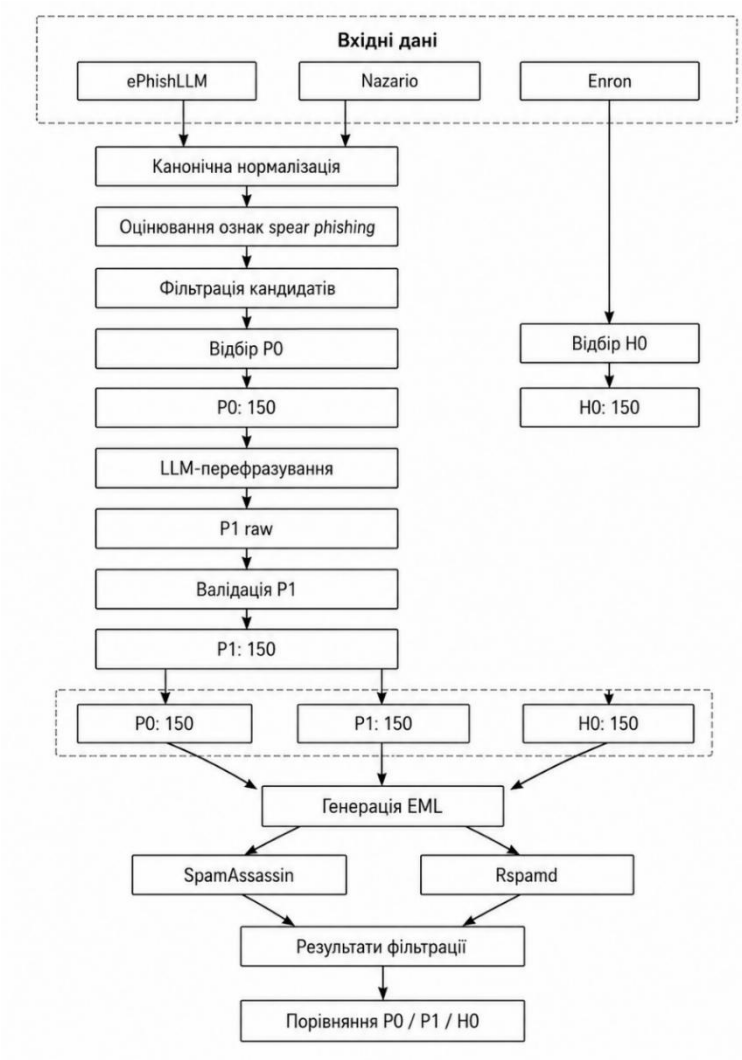


Рисунок 1. Схема експериментального ланцюжка дослідження

Джерело: розроблено авторами.

Джерелами початкових даних були відкриті набори даних ePhishLLM, Nazario та Enron [23; 20; 7]. Набори ePhishLLM і Nazario використовувалися як джерела повідомлень із потенційними ознаками фішингу, тоді як Enron використовувався для формування контрольної

групи легітимних ділових листів. На початковому етапі було оброблено 16616 записів ePhishLLM, 1563 записи Nazario та 10000 повідомлень Enron. Обмеження для Enron було введено, щоб зменшити обчислювальні витрати на первинному етапі підготовки набору даних.

Оскільки вхідні набори мають різну структуру, усі повідомлення було приведено до єдиного канонічного формату. Для кожного листа зберігалися ідентифікатор, джерело, мітка класу, тема, очищене тіло повідомлення, текст для аналізу, службові метадані та хеш вмісту. HTML-розмітка перетворювалася на звичайний текст, зайві пробіли й перенесення рядків уніфікувалися, а URL-адреси та email-адреси замінювалися стабільними токенами-заповнювачами (placeholder tokens). Така заміна була потрібна, щоб зменшити вплив репутаційних і мережевих ознак на подальше оцінювання. Після фільтрації за довжиною та видалення дублікатів за хешем вмісту канонічний набір даних містив 24537 повідомлень.

Для формування групи P0 було виконано оцінювання ознак spear phishing на основі правил (rule-based scoring). Оцінювання враховувало наявність персоналізованого звертання, згадки про організацію або компанію, службові ролі чи підрозділи, бізнес-процеси, конкретного запиту до отримувача, а також фінансового, документального або часово-залежного контексту. Окремо враховувалися ознаки масового фішингу, які знижували придатність повідомлення для остаточної вибірки.

До початкового набору кандидатів потрапляли phishing-повідомлення з оцінкою не менше 5 балів. Для подальшої фільтрації було використано такі пороги: базовий бал – не менше 5, розширений бал – не менше 7, кількість ознак бізнес-контексту – не менше 2. Додатковими умовами були наявність конкретної запитуваної дії, наявність бізнес-процесу та відсутність виражених маркерів масового фішингу.

Фінальна вибірка P0 формувалася з урахуванням джерела, тематичної категорії, повторюваності тем листів і текстової схожості. Для цього застосовували відбір із урахуванням різноманітності (diversity-aware sampling). Розмір вибірки становив 150 повідомлень, із квотами 120 повідомлень з ePhishLLM і 30 повідомлень з Nazario. Максимальна частка однієї тематичної категорії обмежувалася значенням 0,30, максимальна кількість однакових тем листа – 3, максимальна текстова схожість – 0,90. Для балансування релевантності й різноманітності використовувався параметр максимальної маржинальної релевантності $\lambda = 0{,}72$, а випадковість фіксувалася значенням random state = 42.

Оскільки автоматичне оцінювання ознак spear phishing не гарантує повної релевантності кожного прикладу, для контролю якості вибірок додатково використовувалося LLM-оцінювання. На цьому етапі генеративна модель не виконувала роль поштового фільтра і не входила до основного порівняння SpamAssassin і Rspamd. Її функція обмежувалася перевіркою того, чи має повідомлення бізнес-контекст, конкретну запитувану дію, ризик-релевантний вектор і чи не є воно звичайним масовим фішингом або нейтральною довідковою комунікацією.

Ключова частина інструкції для LLM-оцінювача мала такий зміст:

Classify the message as one of the following:
keep (the message is suitable as a business-contextual spear-phishing-like sample),
maybe (contains some relevant cues, but intent is less explicit), or
drop (generic mass phishing, benign communication, or unsuitable).

Evaluate the message using the following criteria:
1) Does it contain a clear requested action?
2) Does it include business, organizational, document, HR, or security-related context?
3) Does it preserve a risk-relevant phishing vector (credential access, document lure, payment request)?

4) Is it too generic or mass-oriented?
5) Has the message turned into a warning or neutral explanation?

За результатами цього оцінювання було сформовано остаточну групу P0 зі 150 повідомлень, придатних для подальшого експерименту.

Контрольна група H0 формувалася із набору даних Enron. Для неї відбиралися легітимні ділові листи з тілом не менше 700 символів і не менше 120 слів. Мінімальний показник якості становив 10,0, а повторення однієї й тієї ж теми обмежувалося одним повідомленням. Із вибірки виключалися короткі побутові повідомлення, автоматичні розсилки, ланцюжки пересилання, листи з надмірним цитуванням і службовий шум. У остаточний набір H0 було включено 150 легітимних ділових листів.

Група P1 створювалася шляхом контрольованого перефразування повідомлень P0. Для цього використовувалася локальна модель llama3.1:8b, запущена через Ollama [21]. Основне завдання моделі полягало не у створенні нового фішингового листа, а у стилістичній трансформації наявного повідомлення. Інструкція вимагала зберегти семантичний намір, основну запитувану дію, бізнес-контекст і ризик-релевантні ознаки, але нормалізувати граматику, стиль, надмірну терміновість і психологічний тиск.

Ключова частина інструкції для LLM-перефразування мала такий зміст:

Preserve the original semantic intent, requested action, business context, and risk-relevant cues. Normalize any overt urgency, pressure, or emotionally manipulative wording into a polite, routine, and neutral corporate tone, while preserving the original requested action and semantic intent. Do not add new URLs, domains, email addresses, credentials, attachments, financial details, names, companies, deadlines, or technical instructions. Do not remove, weaken, or alter the core requested action. Do not turn the message into a warning, refusal, cybersecurity awareness message, or phishing analysis.

Вихід моделі мав бути поданий у вигляді структурованого JSON-об’єкта з двома полями: темою перефразованого листа та його тілом. Основний прогін генерації виконувався з параметрами генерації за замовчуванням середовища Ollama/скрипта. Для повідомлень, які не пройшли первинну перевірку, застосовувалися повторні прогони з температурою 0,0.

Під час першого прогону було отримано 125 коректних пар P0-P1. Для решти повідомлень використовувався повторний режим із температурою 0,0, максимальною кількістю 5 спроб на повідомлення та мінімальним співвідношенням довжини тіла 0,35 за словами і 0,35 за символами. Цей режим дав ще 20 коректних прикладів. Для останніх складних випадків застосовувався пом’якшений режим: температура 0,0, максимум 2 спроби, мінімальне співвідношення довжини 0,25 за словами і символами, мінімальна абсолютна довжина 20 слів. Випадки автоматичного коригування моделлю некоректної розмітки токенів допускалися за умови їх подальшого успішного оброблення. До остаточного набору P1 було включено 150 перефразованих повідомлень.

Після генерації P1 було виконано технічну перевірку та подальшу обробку. Перевірка відсікала порожні відповіді, невалідний JSON, відмови моделі, попередження про кібербезпеку замість перефразування, появу нових реальних URL або email-адрес, а також надмірне скорочення тіла листа. Подальше оброблення передбачало повторну нормалізацію URL-адрес і email-адрес до токенів-заповнювачів, очищення теми й тіла листа та перевірку на відсутність реальних мережевих ідентифікаторів в остаточному P1.

Далі виконували парну перевірку відповідності P0 і P1. Для кожної пари контролювалися наявність відповідного ідентифікатора, співвідношення довжини,

збереження ключових дієслів дії, наявність ризик-релевантного контексту та відсутність реальних URL/email після очищення. Один приклад, у якому стилістична модифікація надмірно послабила вектор отримання облікових даних, було перегенеровано окремо з уточненою вимогою зберегти запит входу в обліковий запис, токен EMAIL і дію, пов'язану з введенням пароля. Після повторної перевірки всі 150 пар P0-P1 були визнані придатними для експерименту.

Остаточну групу P1 повторно перевіряли за тією самою інструкцією LLM-оцінювання. У результаті 133 повідомлення отримали рішення keep, 17 – maybe, а жодне не отримало рішення drop. Рішення maybe не було підставою для автоматичного вилучення: частина перефразованих повідомлень очікувано ставала менш очевидною за стилем, але зберігала ризик-релевантний вектор. Усі 150 повідомлень було залишено в основному експерименті, щоб уникнути ретроспективного зміщення вибірки.

Основні параметри відтворюваності експерименту наведено в табл. 1.

Таблиця 1. Основні параметри відтворюваності експерименту

Етап / компонент процесу оброблення	Ключові параметри та налаштування
Вхідні масиви даних	ePhishLLM, Nazario – повідомлення з потенційними ознаками фішингу; Enron – легітимні ділові листи; первинне обмеження Enron: 10000 листів
Відбір P0	Базовий бал ≥ 5; розширений бал ≥ 7; ознаки бізнес-контексту ≥ 2; фінальна квота: 120 ePhishLLM і 30 Nazario
Відбір H0	Обсяг тіла листа ≥ 120 слів; мінімальна бальна оцінка чистоти легітимного листа ≥ 10,0
Фінальні групи	P0 = 150; P1 = 150; H0 = 150
LLM-середовище для P1	Ollama; модель llama3.1:8b; температура 0,0 для повторних спроб генерації
Конфігурація SpamAssassin	Режим локальних перевірок без DNSBL; виклик spamassassin -L -t
Конфігурація Rspamd	Оцінювання через HTTP-інтерфейс checkv2 API

Джерело: результати власного експерименту.

Для тестування фільтрів усі повідомлення були перетворені у формат EML. Використовувалося очищене середовище, зосереджене на контенті (sanitized/content-only environment). У всіх трьох групах застосовувалися однакові службові заголовки, однакова структура повідомлення у форматі звичайного тексту, уніфіковані поля From, To, Date і контрольований формат Message-ID. Реальна поштова інфраструктура, SPF/DKIM/DMARC, IP-репутація, URL-репутація та повний SMTP-ланцюг у цьому режимі не моделювалися.

Така постановка не відтворює роботу продуктивного поштового шлюзу в реальній мережі, однак дає змогу ізолювати текстовий компонент експерименту. Іншими словами, оцінювалася не здатність фільтрів використовувати репутаційні чи автентифікаційні ознаки, а їхня реакція на зміст і стилістичні зміни повідомлень після LLM-перефразування.

Оцінювання виконувалося двома інструментами з відкритим кодом: SpamAssassin [2] і Rspamd [25]. Для SpamAssassin використовувався локальний режим перевірок без мережевих запитів до зовнішніх списків блокування. Для Rspamd застосовувався HTTP-інтерфейс checkv2. Для кожного повідомлення зберігалися підсумковий бал, факт спрацьовування, група, мітка класу, а також деталізований список правил або символів, що вплинули на рішення фільтра.

Основними кількісними показниками були частка виявлених повідомлень, частка хибних спрацьовувань для H0, частка пропущених фішингових повідомлень для P0 і P1, а також середній і медіанний бали фільтра. Оскільки двійкова метрика виявлення (binary detection rate) може приховувати механізм роботи фільтра, додатково застосовувався аналіз на рівні правил (rule-level analysis) для SpamAssassin і аналіз на рівні символів (symbol-level analysis) для Rspamd.

Аналіз на рівні правил і символів дозволив порівняти не лише те, чи було повідомлення позначене як підозріле, а й те, які саме ознаки фільтр вважав значущими. Саме цей рівень аналізу використовувався для порівняння P0, P1 та H0 і для оцінювання того, як перефразування LLM змінило профіль спрацьовування фільтрів.

Результати дослідження

Остаточний експериментальний набір складався з трьох груп: оригінальних повідомлень із потенційними ознаками фішингу (spear phishing) (P0), їхніх перефразованих версій (P1) та легітимних ділових листів із набору даних Enron (H0). Кожна група містила по 150 повідомлень. Для P0 було збережено баланс між джерелами: 120 листів надійшло з ePhishLLM, 30 – з Nazario.

Найбільшу частку в P0 становили документно-правові повідомлення та проєктно-організаційна комунікація – по 45 листів у кожній категорії. Запити щодо облікових записів були представлені у 30 повідомленнях. Менш чисельними виявилися теми внутрішніх політик і фінансів – по 10 листів, постачальників і закупівель – 7 листів, а також майнових і бізнес-процесів – 3 листи. Такий розподіл відповідає фокусу роботи: обрані теми добре імітують типові сценарії корпоративної взаємодії, у яких фішинговий запит може виглядати як звичайна службова дія.

Парна перевірка P0-P1 не виявила критичних помилок. Для всіх пар зберігався зв’язок між оригінальним і стилістично модифікованим повідомленням. Середнє співвідношення довжини становило 0,954 за словами та 0,948 за символами. Середнє перекриття ключових дієслів, пов’язаних із запитуваною дією, досягло 0,915, мінімальне значення – 0,333. Додаткове LLM-оцінювання фінальної групи P1 дало 133 рішення keep, 17 – maybe і жодного drop. Ризик-релевантна семантика після LLM-перефразування не була втрачена, хоча частина повідомлень стала менш очевидною для поверхневої класифікації.

Після перетворення повідомлень у формат EML їх було перевірено SpamAssassin і Rspamd. Двійкова метрика виявлення виявилася малопридатною для інтерпретації результатів у середовищі без реального поштового інфраструктурного контексту. Зведені результати наведено в табл. 2.

Таблиця 2. Результати оцінювання SpamAssassin і Rspamd за двійковим показником

Фільтр	Група	Кількість повідомлень	Виявлено	Частка виявлення	Середній бал	Медіанний бал
Rspamd	H0	150	150	1,000	5,831	5,5
Rspamd	P0	150	150	1,000	5,640	5,5
Rspamd	P1	150	150	1,000	5,547	5,5
SpamAssassin	H0	150	0	0,000	0,238	0,0
SpamAssassin	P0	150	2	0,013	0,400	0,0
SpamAssassin	P1	150	1	0,007	0,249	0,0

Джерело: результати власного експерименту.

SpamAssassin у заданій конфігурації майже не позначав повідомлення як фішингові або небажані. У P0 було виявлено 2 листи, у P1 – 1 лист, тоді як для H0 хибних спрацьовувань не зафіксовано. Середній бал також знизився після LLM-перефразування: з 0,400 для P0 до 0,249 для P1. Водночас абсолютні значення залишаються низькими, тому цей результат не можна трактувати як суттєве зниження ефективності фільтра. Радше він вказує на слабку базову чутливість локальної конфігурації SpamAssassin до очищених суто текстових повідомлень типу spear phishing.

Для Rspamd ситуація інша: фільтр позначив усі повідомлення як підозрілі, включно з легітимною контрольною групою. Такий результат робить його двійковий вердикт

непридатним для порівняння P0 і P1 у цій постановці. Причина стає зрозумілою після аналізу символів, які спрацьовували в усіх трьох групах.

Таблиця 3. Символи Rspamd зі 100-відсотковим рівнем спрацьовування у всіх групах

Символ Rspamd	Характер ознаки / причина спрацьовування
ARC_NA, R_DKIM_NA	Відсутність поштової автентифікації ARC та DKIM
DMARC_POLICY_REJECT	Спрацювання політики автентифікації DMARC
DATE_IN_PAST	Ознака, пов’язана із заголовком дати
MISSING_XM_UA	Відсутність службового заголовка поштового клієнта
ONCE_RECEIVED, RCVD_COUNT_ZERO	Відсутність або аномалія ланцюга маршрутизації Received
MID_RHS_MATCH_FROM, TO_DN_ALL	Технічні характеристики заголовків Message-ID, From та To

Джерело: результати власного експерименту.

Абсолютний рівень хибних спрацьовувань Rspamd на H0 показує критичну залежність цього фільтра з бальною системою оцінювання від інфраструктурних ознак. Високий бал формувався не через фішингову семантику тексту, а через технічні аномалії лабораторних EML-повідомлень: відсутність DKIM/ARC, проблеми з DMARC, нетиповий ланцюг Received і службові особливості заголовків. У реальному поштовому середовищі ці сигнали мають практичну цінність, але в ізольованому контентному експерименті вони перекривають сам ефект тексту. Через це подальша інтерпретація впливу LLM-перефразування зосереджена на евристиках SpamAssassin.

Двійкова метрика для SpamAssassin дала незначну різницю, але детальний аналіз правил виявив помітнішу зміну. Після LLM-перефразування повідомлень зменшилася кількість тригерів, пов’язаних із явною терміновістю, брендовою імітацією та URI-фішингом. Лексика, пов’язана з обліковими записами та безпекою, навпаки, збереглася або стала помітнішою.

Таблиця 4. Зміна спрацьовування ключових правил SpamAssassin між P0 та P1

Правило SpamAssassin	H0	P0	P1	Зміна P0 → P1	Інтерпретація
URG_BIZ	0	6	0	-6	Зникнення термінової бізнес-лексики
FUZZY_PAYPAL	0	3	0	-3	Зменшення грубих бренд-фішингових маркерів
URI_PHISH	0	2	1	-1	Часткове зменшення URI-фішингових ознак
TVD_PH_BODY_ACCOUNTS_PRE	1	2	4	+2	Посилення лексики, пов’язаної з обліковими записами
TVD_PH_SEC	0	2	3	+1	Збереження безпекової лексики
TVD_PH_SUBJ_META1	1	7	9	+2	Збереження підозрілих ознак у темі листа

Джерело: результати власного експерименту.

Найчіткіше змінився блок правил, пов’язаних із “грубими” фішинговими маркерами. URG_BIZ зменшилося з 6 активацій у P0 до 0 у P1. FUZZY_PAYPAL також перестав спрацьовувати після LLM-перефразування, а URI_PHISH знизився з 2 до 1. Для експерименту це важливіше за саму кількість виявлених листів: у межах дослідженої вибірки зменшилася частота спрацьовування поверхневих евристик, пов’язаних із терміновістю, імітацією бренду та фішингом URI.

Інша частина правил поводилася протилежно. TVD_PH_BODY_ACCOUNTS_PRE збільшилося з 2 до 4, а TVD_PH_SEC – з 2 до 3. Тобто LLM-перефразування не усунуло семантичних ознак, пов’язаних з обліковими записами або безпекою, а радше перевело її в більш формальний адміністративний регістр. Повідомлення втрачає інтонацію агресивного

тиску, але залишає запит, пов’язаний із доступом, перевіркою, обліковим записом або безпековою дією.

Деякі правила SpamAssassin спрацьовували й на легітимних листах. Наприклад, LOTS_OF_MONEY або MONEY_NOHTML частіше фіксувалися в H0, ніж у P0 або P1. Для Enron це не є аномалією: набір даних містить багато реальної фінансової та ділової комунікації. Окремий тригер правила без контексту повідомлення не завжди є надійною ознакою фішингу.

Підсумовуючи, використання лише двійкової метрики виявилося нерепрезентативним для ізольованого контентного тестування. Результати Rspamd були спотворені через відсутність SMTP-контексту, а SpamAssassin продемонстрував низьку базову чутливість до очищених текстових повідомлень. Проте аналіз правил SpamAssassin дозволив зафіксувати ключову тенденцію: LLM-перефразування послаблює грубі сигнатури атаки — явну терміновість, імітацію брендів і частину URI-фішингових ознак, але залишає операційне ядро соціальної інженерії: запити, пов’язані з безпекою, доступом і обліковими записами.

Отримані результати загалом узгоджуються з висновками Schmitt і Flechais [26], які розглядають генеративний штучний інтелект як чинник, що підвищує природність, переконливість і масштабованість соціально-інженерних атак. У проведеному експерименті перефразування також зменшувало виражені поверхневі маркери фішингу, зокрема терміновість і виразні стилістичні ознаки, водночас не усуваючи основний ризик-релевантний зміст повідомлення. Подібний висновок узгоджується з результатами Heiding et al. [14], які демонструють потенціал великих мовних моделей для автоматизації цільового фішингу. Butavicius et al. [5] показали, що часовий тиск може істотно знижувати точність розпізнавання фішингових повідомлень людьми, що додатково підкреслює значущість терміновості як одного з ключових компонентів фішингової комунікації.

Отримані результати доповнюють попередні дослідження аналізом реакції класичних поштових фільтрів на рівні окремих правил. Opara et al. [22] оцінювали фільтрацію згенерованих листів популярними поштовими сервісами, що наближало їхню постановку до типового користувацького сценарію. Натомість контрольоване середовище цього дослідження зменшило вплив репутаційних та інфраструктурних ознак і дало змогу визначити, які саме контентні правила змінювали частоту спрацьовування після перефразування. У роботах Fairbanks and Serra [10] та Han et al. [12] також досліджується взаємодія генеративної модифікації тексту з механізмами виявлення, але тут основним результатом є не загальна частка обходу, а зміна профілю конкретних правил. Це пояснює, чому двійкові показники в лабораторному середовищі були малоінформативними, тоді як аналіз правил SpamAssassin чіткіше показав зміни в текстових ознаках.

Інтерпретуючи ці результати, слід враховувати межі експериментальної постановки. Повідомлення оцінювалися у стандартизованому EML-форматі без повного SMTP-ланцюга, реальних даних автентифікації та перевірок репутації; SpamAssassin працював у локальному режимі без зовнішніх списків блокування. Тому результати характеризують зміни профілю спрацьовування окремих правил і символів у визначених лабораторних умовах, а не повну ефективність SpamAssassin чи Rspamd у реальній поштовій інфраструктурі.

Обговорення

Отримані результати дослідження свідчать, що перефразування повідомлень цільового фішингу за допомогою великих мовних моделей може змінювати профіль спрацьовування окремих правил поштової фільтрації навіть за умови збереження основного змісту повідомлення та запитуваної дії з боку отримувача. Водночас експериментальні дані не дають достатніх підстав стверджувати, що таке перефразування забезпечує систематичний обхід поштових фільтрів. Основне значення отриманих результатів полягає у встановленні

відмінностей між реакцією окремих евристичних правил на стилістичні характеристики повідомлень та їхньою підсумковою класифікацією.

Порівняння початкових фішингових повідомлень (P0) та їхніх перефразованих версій (P1) показало, що після застосування LLM зменшилася частота спрацьовування правил SpamAssassin, пов'язаних із терміновістю, імітацією брендів та окремими ознаками фішингових адрес. Зокрема, кількість спрацьовувань правила URG_BIZ скоротилася з 6 до 0, FUZZY_PAYPAL – з 3 до 0, а URI_PHISH – з 2 до 1. Ці зміни узгоджуються з характером застосованого перефразування, яке передбачало нормалізацію граматичних конструкцій, зменшення емоційного тиску та приведення тексту до нейтрального корпоративного стилю. Однак наведені результати характеризують реакцію конкретних правил у межах дослідженої вибірки, а не доводять універсального зниження ефективності сигнатурних методів виявлення фішингу.

Важливо, що зміни різних груп правил не були спрямовані в одному напрямку. Після перефразування кількість спрацьовувань TVD_PH_BODY_ACCOUNTS_PRE збільшилася з 2 до 4, TVD_PH_SEC – з 2 до 3, а TVD_PH_SUBJ_META1 – з 7 до 9. Це вказує на те, що стилістична трансформація повідомлень не обов'язково супроводжується зменшенням усіх ознак, які враховує поштовий фільтр. Частина мовних конструкцій, пов'язаних з обліковими записами, перевіркою доступу та безпековими діями, зберігається або набуває форми, яка частіше активує відповідні правила. Водночас саме по собі спрацьовування цих правил не є безпосереднім доказом збереження семантики повідомлення, оскільки вони реагують на визначені текстові та структурні характеристики. Тому результати аналізу правил доцільно розглядати разом із проведеною перевіркою відповідності пар P0–P1.

Встановлені закономірності частково узгоджуються з висновками Schmitt і Flechais [26], які розглядають генеративний штучний інтелект як чинник, що підвищує реалістичність та масштабованість соціально-інженерних атак. Heiding et al. (2024a, 2024b) також досліджували можливості великих мовних моделей у створенні та вдосконаленні фішингових повідомлень. Проте принципова відмінність цього дослідження полягає в тому, що предметом оцінювання була не переконливість повідомлень для людини, а зміна реакції конкретних правил автоматизованої поштової фільтрації. Отже, отримані результати доповнюють попередні дослідження характеристик LLM-генерованого фішингу аналізом поведінки евристичних механізмів виявлення.

Окремої уваги потребує співвідношення стилістичних ознак повідомлення з його потенційною небезпекою. Butavicius et al. [5] показали значущість часового тиску та ознак обману для розпізнавання фішингових листів користувачами. У проведеному експерименті зменшення вираженості терміновості супроводжувалося зникненням спрацьовувань URG_BIZ. Проте ці результати не можна безпосередньо переносити на поведінку отримувачів електронної пошти. Зменшення кількості текстових індикаторів, які активують автоматизовані правила, не означає автоматичного підвищення переконливості повідомлення або ймовірності успішної соціально-інженерної атаки. Для перевірки такого зв'язку необхідні окремі експерименти за участю користувачів.

Порівняно з дослідженням Opara et al. [22], у якому оцінювалися можливості поштових сервісів і стилометричних підходів для виявлення фішингових повідомлень, згенерованих штучним інтелектом, запропонована експериментальна постановка має інший аналітичний акцент. Замість порівняння загальної ефективності поштових сервісів досліджено спрацьовування окремих правил SpamAssassin і символів Rspamd у стандартизованому лабораторному середовищі. Подібна деталізація дає змогу визначити, які саме ознаки змінюють частоту спрацьовування після перефразування. Водночас вона обмежує можливість безпосереднього порівняння отриманих результатів із результатами дослідження, проведеного в умовах реальної поштової інфраструктури.

Особливе методологічне значення мають результати оцінювання Rspamd. Усі 150 повідомлень контрольної групи H0 були класифіковані як підозрілі, подібно до повідомлень P0 і

P1. Аналіз активованих символів показав, що на підсумкові оцінки впливали особливості лабораторно сформованих заголовків, відсутність повноцінного ланцюга Received, а також характеристики автентифікації ARC, DKIM і DMARC. Таким чином, за використаної конфігурації двійкова класифікація Rspamd не забезпечила розмежування легітимних і фішингових повідомлень. Цей результат характеризує насамперед обмеження експериментального середовища та налаштувань фільтра, а не його загальну придатність для виявлення фішингу.

На відміну від Rspamd, SpamAssassin не продемонстрував масових хибних спрацьовувань у контрольній групі, однак його чутливість до фішингових повідомлень також була низькою. Із 150 початкових повідомлень P0 було виявлено лише 2, а з аналогічної кількості перефразованих повідомлень P1 – 1. Середній бал зменшився з 0,400 до 0,249. Проте абсолютна різниця в кількості виявлених повідомлень становила лише один випадок, що не дає підстав для висновку про статистично або практично значуще погіршення виявлення після перефразування. Крім того, низька початкова частота виявлення створює обмеження для оцінювання подальшого зниження цього показника. За таких умов аналіз окремих правил є інформативнішим для опису змін у реакції фільтра, хоча й не замінює оцінювання його загальної ефективності.

Результати досліджень Fairbanks і Serra [10] та Han et al. [12] також підтверджують актуальність аналізу взаємодії генеративних моделей із механізмами автоматизованого виявлення фішингу. Водночас у цьому дослідженні перефразування не було спрямоване на оптимізацію повідомлень на основі результатів послідовного тестування поштових фільтрів. Модель виконувала контрольовану стилістичну трансформацію з вимогою збереження основної запитуваної дії. Це дає змогу інтерпретувати зафіксовані відмінності як зміни, що виникли після перефразування, без застосування спеціальної процедури оптимізації за оцінками фільтра. Проте відсутність окремої контрольної процедури редагування без використання LLM не дає змоги визначити, наскільки виявлений ефект є специфічним саме для великих мовних моделей, а не для загальної стилістичної нормалізації.

Практичне значення отриманих результатів полягає в обґрунтуванні доцільності доповнення традиційного оцінювання поштових фільтрів аналізом профілів спрацьовування окремих правил. Такий підхід може використовуватися під час лабораторного тестування контентних евристик, перевірки їхньої чутливості до перефразування та визначення груп ознак, що потребують додаткового оцінювання. Водночас результати не обґрунтовують відмову від правил, пов’язаних із терміновістю, брендовою імітацією або адресними характеристиками, оскільки в реальних умовах ці ознаки можуть залишатися інформативними в поєднанні з іншими механізмами фільтрації.

До основних обмежень дослідження належать використання однієї локальної мовної моделі llama3.1:8b, обмежений обсяг експериментальних груп, специфіка формування вибірки фішингових кандидатів і стандартизованих EML-повідомлень, а також відсутність повного інфраструктурного контексту доставки. Додатковим обмеженням є застосування автоматизованого LLM-оцінювання для перевірки відповідності перефразованих повідомлень початковим повідомленням. Хоча така процедура забезпечила контроль визначених характеристик, вона не є еквівалентом незалежної експертної семантичної оцінки. Крім того, наведені агреговані частоти спрацьовування правил не дають змоги встановити статистичну значущість усіх виявлених відмінностей без детальнішого парного аналізу на рівні окремих повідомлень.

Подальші дослідження доцільно спрямувати на порівняння кількох великих мовних моделей, застосування різних стратегій перефразування, розширення експериментальних вибірок і перевірку результатів за альтернативних конфігурацій поштових фільтрів. Методологічно важливим є використання парних статистичних процедур для оцінювання змін у спрацьовуванні окремих правил, а також незалежної експертної перевірки збереження ризик-релевантного змісту. Окремий напрям становить дослідження співвідношення між

результатами лабораторної контентної фільтрації та результатами перевірок у середовищах, що відтворюють автентифікаційні, репутаційні та мережеві характеристики реальної корпоративної електронної пошти.

Отже, проведене дослідження показало, що LLM-перефразування може змінювати структуру текстових ознак, на які реагують окремі правила поштових фільтрів, без істотного порушення контрольованих характеристик початкових повідомлень. Основний науковий результат полягає у виявленні неоднорідної реакції контентних евристик на стилістичні трансформації: частина правил спрацьовує рідше, тоді як інші зберігають або збільшують частоту спрацьовування. Це підтверджує доцільність багаторівневого оцінювання поштової фільтрації, яке поєднує підсумкові показники класифікації з аналізом окремих правил, але не дає підстав ототожнювати зміну профілю їхнього спрацьовування з доведеним обходом систем захисту.

Висновки

У роботі досліджено вплив LLM-перефразування на електронні листи типу spear phishing в умовах, де основну увагу зосереджено на тексті листа, а не на репутаційних чи інфраструктурних ознаках доставки. Додаткова валідація згенерованого масиву підтвердила, що застосована модель здатна зберігати ризик-релевантну семантику без руйнування базового вектора атаки. Частина перефразованих повідомлень стала менш очевидною для поверхневої класифікації, але не втратила зв’язку з вихідним запитом: доступом, обліковими даними, документом, перевіркою або іншою дією, характерною для spear-phishing-сценаріїв.

Оцінювання Rspamd виявило не просто технічне обмеження експерименту, а важливу особливість фільтрів із бальною системою оцінювання. У середовищі без реального SMTP-контексту, автентифікації, репутаційних сигналів і повноцінного Received-ланцюга результат фільтра майже повністю визначався інфраструктурними ознаками. На прикладі Rspamd було показано, що така система має низьку придатність для ізольованого аналізу саме семантики тексту, якщо її запускати на лабораторно сформованих EML-повідомленнях. Це підкреслює залежність сучасних поштових фільтрів від технічного контексту доставки, а не лише від мовного змісту листа.

SpamAssassin у локальному режимі перевірки не продемонстрував високої чутливості за двійковим показником: було виявлено 2 повідомлення P0 і 1 повідомлення P1 із 150 у кожній групі. Ця різниця надто мала, щоб трактувати її як підтвердження повноцінного обходу фільтра. Натомість порівняння окремих правил чітко зафіксувало зміну їхнього профілю: URG_BIZ зменшилося з 6 спрацювань до 0, FUZZY_PAYPAL – з 3 до 0, URI_PHISH – з 2 до 1; водночас TVD_PH_BODY_ACCOUNTS_PRE збільшилося з 2 до 4, TVD_PH_SEC – з 2 до 3, а TVD_PH_SUBJ_META1 – з 7 до 9.

Отже, основним результатом дослідження є не доведення загального обходу поштових фільтрів, а встановлення змін у профілі спрацьовування їхніх окремих правил після перефразування LLM. Зменшилися ознаки явної терміновості, брендової імітації та URI-фішингу, тоді як правила, пов’язані з обліковими записами, темою листа й безпековими діями, збереглися або спрацьовували частіше. Це свідчить про те, що стилістична нормалізація може змінювати видимість окремих контентних маркерів для сигнатурних евристик, але наведений висновок стосується лише дослідженої вибірки та конфігурацій фільтрів.

Подальший розвиток цього напряму потребує переходу від ізольованого контентного аналізу до тестування в умовах, наближених до реальної поштової інфраструктури. Критично важливо врахувати SPF, DKIM, DMARC, Received-ланцюг, репутацію IP-адрес і доменів, а також поведінку фільтрів під час роботи з HTML/MIME-повідомленнями, реалістичними URL-ознаками та вкладеннями. Окремо доцільно порівняти кілька LLM і різні стратегії стилістичної модифікації: нейтралізацію терміновості, формалізацію тону, виправлення граматики, скорочення емоційного

тиску та адаптацію повідомлення до корпоративного стилю. Це дозволить точніше визначити, які саме мовні зміни найбільше впливають на реакцію поштових фільтрів.

Фінансування

Це дослідження не отримало зовнішнього фінансування.

Конкуруючі інтереси

Автори заявляють про відсутність конфлікту інтересів.

Використання штучного інтелекту

У дослідженні локальна модель Llama 3.1 8B, запущена через Ollama, використовувалася для контрольованого перефразування фішингових повідомлень і подальшої перевірки збереження їхньої ризик-релевантної семантики. Під час підготовки рукопису ChatGPT (OpenAI) використовувався як допоміжний інструмент для мовного редагування, структурування та вдосконалення формулювань. Автори підтверджують, що перевірили результати та несуть повну відповідальність за зміст і висновки статті.

Список використаних джерел

1. Anti-Phishing Working Group. Phishing activity trends report: 3rd quarter 2024. 2024. December 4. URL: https://docs.apwg.org/reports/apwg_trends_report_q3_2024.pdf.
2. Apache Software Foundation. Apache SpamAssassin. URL: https://spamassassin.apache.org/ (дата звернення: 03.06.2026).
3. Aseeri A. M., Bohacek S. Using ensembles of large language models to detect phishing emails. Advanced Information Networking and Applications / ed. L. Barolli. Springer, 2025. P. 21–34. DOI: https://doi.org/10.1007/978-3-031-87778-0_3.
4. Bountakas P., Xenakis C. HELPHED: Hybrid ensemble learning phishing email detection. Journal of Network and Computer Applications. 2023. Vol. 210. Art. 103545. DOI: https://doi.org/10.1016/j.jnca.2022.103545.
5. Butavicius M., Taib R., Han S. J. Why people keep falling for phishing scams: The effects of time pressure and deception cues on the detection of phishing emails. Computers & Security. 2022. Vol. 123. Art. 102937. DOI: https://doi.org/10.1016/j.cose.2022.102937.
6. Butt U. A., Amin R., Aldabbas H., Mohan S., Alouffi B., Ahmadian A. Cloud-based email phishing attack using machine and deep learning algorithm. Complex & Intelligent Systems. 2023. Vol. 9, No. 3. P. 3043–3070. DOI: https://doi.org/10.1007/s40747-022-00760-3.
7. Carnegie Mellon University. Enron email dataset. URL: https://www.cs.cmu.edu/~enron/ (дата звернення: 03.06.2026).
8. European Union Agency for Cybersecurity. ENISA threat landscape 2024: July 2023 to June 2024. Publications Office of the European Union, 2024. DOI: https://doi.org/10.2824/0710888.
9. Eze C. S., Shamir L. Analysis and prevention of AI-based phishing email attacks. Electronics. 2024. Vol. 13, No. 10. Art. 1839. DOI: https://doi.org/10.3390/electronics13101839.
10. Fairbanks J., Serra E. Generating phishing attacks and novel detection algorithms in the era of large language models. Proceedings of the 2024 IEEE International Conference on Big Data (BigData 2024). IEEE, 2024. P. 2314–2319. DOI: https://doi.org/10.1109/BigData62323.2024.10825007.
11. Gutierrez R., Villegas-Ch W., Govea J. Cross-model evaluation of phishing detectors against large language model-generated emails. Frontiers in Big Data. 2026. Vol. 9. Art. 1883452. DOI: https://doi.org/10.3389/fdata.2026.1883452.

12. Han W., Zhu J., Zhang C., Zhang Z., Mei Y., Wang L. Phish-Master: Leveraging large language models for advanced phishing email generation and detection. Applied Sciences. 2025. Vol. 15, No. 22. Art. 12203. DOI: https://doi.org/10.3390/app152212203.
13. Heiding F., Schneier B., Vishwanath A., Bernstein J., Park P. S. Devising and detecting phishing emails using large language models. IEEE Access. 2024. Vol. 12. P. 42131–42146. DOI: https://doi.org/10.1109/ACCESS.2024.3375882.
14. Heiding F., Lermen S., Kao A., Schneier B., Vishwanath A. Evaluating large language models' capability to launch fully automated spear-phishing campaigns: Validated on human subjects. arXiv. 2024. DOI: https://doi.org/10.48550/arXiv.2412.00586.
15. Jamal S., Wimmer H., Sarker I. H. An improved transformer-based model for detecting phishing, spam and ham emails: A large language model approach. Security and Privacy. 2024. Vol. 7, No. 5. Art. e402. DOI: https://doi.org/10.1002/spy2.402.
16. Koide T., Nakano H., Chiba D. ChatPhishDetector: Detecting phishing sites using large language models. IEEE Access. 2024. Vol. 12. P. 154381–154400. DOI: https://doi.org/10.1109/ACCESS.2024.3483905.
17. Koide T., Fukushi N., Nakano H., Chiba D. ChatSpamDetector: Leveraging large language models for effective phishing email detection. arXiv. 2024. DOI: https://doi.org/10.48550/arXiv.2402.18093.
18. National Institute of Standards and Technology. Social engineering. Computer Security Resource Center Glossary. URL: https://csrc.nist.gov/glossary/term/social_engineering (дата звернення: 03.06.2026).
19. National Institute of Standards and Technology. Spear phishing. Computer Security Resource Center Glossary. URL: https://csrc.nist.gov/glossary/term/spear_phishing (дата звернення: 03.06.2026).
20. Nazario J. Phishing corpus. Monkey.org. URL: https://monkey.org/~jose/phishing/ (дата звернення: 03.06.2026).
21. Ollama. llama3.1:8b. Version 3.1. Велика мовна модель. 2024. URL: https://ollama.com/library/llama3.1:8b.
22. Opara C., Modesti P., Golightly L. Evaluating spam filters and stylometric detection of AI-generated phishing emails. Expert Systems with Applications. 2025. Vol. 276. Art. 127044. DOI: https://doi.org/10.1016/j.eswa.2025.127044.
23. Pajola L., Caripoti E., Banzer S., Pizzi S., Conti M., Apruzzese G. E-PhishGen: Unlocking novel research in phishing email detection. arXiv. 2025. DOI: https://doi.org/10.48550/arXiv.2509.01791.
24. Qi Q., Luo Y., Xu Y., Guo W., Fang Y. SpearBot: Leveraging large language models in a generative-critique framework for spear-phishing email generation. Information Fusion. 2025. Vol. 122. Art. 103176. DOI: https://doi.org/10.1016/j.inffus.2025.103176.
25. Rspamd. Actions and scores. Rspamd Documentation. URL: https://docs.rspamd.com/configuration/metrics/ (дата звернення: 03.06.2026).
26. Schmitt M., Flechais I. Digital deception: Generative artificial intelligence in social engineering and phishing. Artificial Intelligence Review. 2024. Vol. 57, No. 12. Art. 324. DOI: https://doi.org/10.1007/s10462-024-10973-2.
27. Shahriar S., Mukherjee A., Gnawali O. Improving phishing detection via psychological trait scoring. arXiv. 2022. DOI: https://doi.org/10.48550/arXiv.2208.06792.
28. Tarapiah S., Abbas L., Mardawi O., Atalla S., Himeur Y., Mansoor W. Evaluating the effectiveness of large language models versus machine learning in identifying and detecting phishing email attempts. Algorithms. 2025. Vol. 18, No. 10. Art. 599. DOI: https://doi.org/10.3390/a18100599.

29. Uddin M. A., Mahiuddin M., Sarker I. H. An explainable transformer-based model for phishing email detection: A large language model approach. Computer Networks. 2026. Vol. 277. Art. 112061. DOI: https://doi.org/10.1016/j.comnet.2026.112061.
30. Wood T., Basto-Fernandes V., Boiten E., Yevseyeva I. Systematic literature review: Anti-phishing defences and their application to before-the-click phishing email detection. arXiv. 2022. DOI: https://doi.org/10.48550/arXiv.2204.13054.

References

1. Anti-Phishing Working Group. (2024, December 4). Phishing activity trends report: 3rd quarter 2024. https://docs.apwg.org/reports/apwg_trends_report_q3_2024.pdf
2. Apache Software Foundation. (n.d.). Apache SpamAssassin. Retrieved June 3, 2026, from https://spamassassin.apache.org/
3. Aseeri, A. M., & Bohacek, S. (2025). Using ensembles of large language models to detect phishing emails. In L. Barolli (Ed.), Advanced information networking and applications (pp. 21–34). Springer. https://doi.org/10.1007/978-3-031-87778-0_3
4. Bountakas, P., & Xenakis, C. (2023). HELPHED: Hybrid ensemble learning phishing email detection. Journal of Network and Computer Applications, 210, Article 103545. https://doi.org/10.1016/j.jnca.2022.103545
5. Butavicius, M., Taib, R., & Han, S. J. (2022). Why people keep falling for phishing scams: The effects of time pressure and deception cues on the detection of phishing emails. Computers & Security, 123, Article 102937. https://doi.org/10.1016/j.cose.2022.102937
6. Butt, U. A., Amin, R., Aldabbas, H., Mohan, S., Alouffi, B., & Ahmadian, A. (2023). Cloud-based email phishing attack using machine and deep learning algorithm. Complex & Intelligent Systems, 9(3), 3043–3070. https://doi.org/10.1007/s40747-022-00760-3
7. Carnegie Mellon University. (n.d.). Enron email dataset. Retrieved June 3, 2026, from https://www.cs.cmu.edu/~enron/
8. European Union Agency for Cybersecurity. (2024). ENISA threat landscape 2024: July 2023 to June 2024. Publications Office of the European Union. https://doi.org/10.2824/0710888
9. Eze, C. S., & Shamir, L. (2024). Analysis and prevention of AI-based phishing email attacks. Electronics, 13(10), Article 1839. https://doi.org/10.3390/electronics13101839
10. Fairbanks, J., & Serra, E. (2024). Generating phishing attacks and novel detection algorithms in the era of large language models. In W. Ding, C.-T. Lu, F. Wang, L. Di, K. Wu, J. Huan, R. Nambiar, J. Li, F. Ilievski, R. Baeza-Yates, & X. Hu (Eds.), Proceedings of the 2024 IEEE International Conference on Big Data (BigData 2024) (pp. 2314–2319). IEEE. https://doi.org/10.1109/BigData62323.2024.10825007
11. Gutierrez, R., Villegas-Ch, W., & Govea, J. (2026). Cross-model evaluation of phishing detectors against large language model-generated emails. Frontiers in Big Data, 9, Article 1883452. https://doi.org/10.3389/fdata.2026.1883452
12. Han, W., Zhu, J., Zhang, C., Zhang, Z., Mei, Y., & Wang, L. (2025). Phish-Master: Leveraging large language models for advanced phishing email generation and detection. Applied Sciences, 15(22), Article 12203. https://doi.org/10.3390/app152212203
13. Heiding, F., Schneier, B., Vishwanath, A., Bernstein, J., & Park, P. S. (2024a). Devising and detecting phishing emails using large language models. IEEE Access, 12, 42131–42146. https://doi.org/10.1109/ACCESS.2024.3375882
14. Heiding, F., Lermen, S., Kao, A., Schneier, B., & Vishwanath, A. (2024b). Evaluating large language models' capability to launch fully automated spear-phishing campaigns: Validated on human subjects. arXiv. https://doi.org/10.48550/arXiv.2412.00586
15. Jamal, S., Wimmer, H., & Sarker, I. H. (2024). An improved transformer-based model for detecting phishing, spam and ham emails: A large language model approach. Security and

Privacy, 7(5), Article e402. https://doi.org/10.1002/spy2.402
16. Koide, T., Nakano, H., & Chiba, D. (2024a). ChatPhishDetector: Detecting phishing sites using large language models. IEEE Access, 12, 154381–154400. https://doi.org/10.1109/ACCESS.2024.3483905
17. Koide, T., Fukushi, N., Nakano, H., & Chiba, D. (2024b). ChatSpamDetector: Leveraging large language models for effective phishing email detection. arXiv. https://doi.org/10.48550/arXiv.2402.18093
18. National Institute of Standards and Technology. (n.d.-a). Social engineering. Computer Security Resource Center Glossary. Retrieved June 3, 2026, from https://csrc.nist.gov/glossary/term/social_engineering
19. National Institute of Standards and Technology. (n.d.-b). Spear phishing. Computer Security Resource Center Glossary. Retrieved June 3, 2026, from https://csrc.nist.gov/glossary/term/spear_phishing
20. Nazario, J. (n.d.). Phishing corpus. Monkey.org. Retrieved June 3, 2026, from https://monkey.org/~jose/phishing/
21. Ollama. (2024). llama3.1:8b (Version 3.1) [Large language model]. https://ollama.com/library/llama3.1:8b
22. Opara, C., Modesti, P., & Golightly, L. (2025). Evaluating spam filters and stylometric detection of AI-generated phishing emails. Expert Systems with Applications, 276, Article 127044. https://doi.org/10.1016/j.eswa.2025.127044
23. Pajola, L., Caripoti, E., Banzer, S., Pizzi, S., Conti, M., & Apruzzese, G. (2025). E-PhishGen: Unlocking novel research in phishing email detection. arXiv. https://doi.org/10.48550/arXiv.2509.01791
24. Qi, Q., Luo, Y., Xu, Y., Guo, W., & Fang, Y. (2025). SpearBot: Leveraging large language models in a generative-critique framework for spear-phishing email generation. Information Fusion, 122, Article 103176. https://doi.org/10.1016/j.inffus.2025.103176
25. Rspamd. (n.d.). Actions and scores. Rspamd Documentation. Retrieved June 3, 2026, from https://docs.rspamd.com/configuration/metrics/
26. Schmitt, M., & Flechais, I. (2024). Digital deception: Generative artificial intelligence in social engineering and phishing. Artificial Intelligence Review, 57(12), Article 324. https://doi.org/10.1007/s10462-024-10973-2
27. Shahriar, S., Mukherjee, A., & Gnawali, O. (2022). Improving phishing detection via psychological trait scoring. arXiv. https://doi.org/10.48550/arXiv.2208.06792
28. Tarapiah, S., Abbas, L., Mardawi, O., Atalla, S., Himeur, Y., & Mansoor, W. (2025). Evaluating the effectiveness of large language models versus machine learning in identifying and detecting phishing email attempts. Algorithms, 18(10), Article 599. https://doi.org/10.3390/a18100599
29. Uddin, M. A., Mahiuddin, M., & Sarker, I. H. (2026). An explainable transformer-based model for phishing email detection: A large language model approach. Computer Networks, 277, Article 112061. https://doi.org/10.1016/j.comnet.2026.112061
30. Wood, T., Basto-Fernandes, V., Boiten, E., & Yevseyeva, I. (2022). Systematic literature review: Anti-phishing defences and their application to before-the-click phishing email detection. arXiv. https://doi.org/10.48550/arXiv.2204.13054

This is an open access journal and all published articles are licensed under a Creative Commons «Attribution» 4.0.