КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 2 (34), 2026
ISSN 2663 – 4023
DOI 10.28925/2663-4023.2026.34.1269
УДК 004.056.5:004.8

Куцюк Даниїл Андрійович
студент бакалаврської програми «Аналіз вразливостей інформаційних систем»
Національний університет «Києво-Могилянська академія», Київ, Україна
ORCID:0009-0008-1018-3767
d.kutsiuk@ukma.edu.ua

Гороховський Кирило Семенович
старший викладач кафедри мультимедійних систем факультету інформатики
Національний університет «Києво-Могилянська академія», Київ, Україна
ORCID:0000-0002-6398-5367
fintech.lab@ukma.edu.ua

РОЗРОБКА МОДУЛЯ AI-АСИСТЕНТА НА ОСНОВІ ДОНАВЧАННЯ ВЕЛИКОЇ МОВНОЇ МОДЕЛІ МЕТОДОМ QLoRA ДЛЯ АВТОМАТИЗОВАНОГО АНАЛІЗУ ІНЦИДЕНТІВ КІБЕРБЕЗПЕКИ

Анотація. У статті розглянуто розробку та інтеграцію модуля штучного інтелекту для автоматизованого аналізу інцидентів кібербезпеки в межах децентралізованої платформи обміну даними про кіберінциденти Cybersecurity Dataspace (CSDS). Актуальність дослідження зумовлена системними проблемами сучасних центрів управління безпекою (SOC): надмірним обсягом подій безпеки, глобальним дефіцитом кваліфікованих фахівців, ефектом «втоми від алертів» та зростанням складності багатоетапних атак. Метою роботи є побудова спеціалізованої великої мовної моделі (LLM), здатної класифікувати загрози, оцінювати рівень їх серйозності та надавати структуровані рекомендації щодо реагування відповідно до методологій MITRE ATT&CK та NIST SP 800-61. У роботі виконано порівняльний аналіз чотирьох підходів до побудови AI-асистентів (zero-shot промптинг, RAG, повне донавчання та параметрично-ефективне донавчання) і обґрунтовано вибір методу QLoRA як оптимального для локального розгортання з огляду на вимоги конфіденційності та обмежені обчислювальні ресурси. Сформовано навчальний датасет обсягом 1875 прикладів із п’яти відкритих джерел (MITRE ATT&CK, NVD CVE, синтетичні логи інцидентів, відкриті Q&A-датасети та правила виявлення Sigma/Elastic), який пройшов дедуплікацію, фільтрацію за якістю та верифікацію технік MITRE. Проведено два незалежні раунди донавчання базової моделі Llama 3.2 3B Instruct методом QLoRA з ідентичними гіперпараметрами; описано їхні результати за метриками функції втрат і точності на рівні токенів. Розроблено мікросервіс на базі FastAPI з REST-ендпоінтами для аналізу інцидентів, діалогу, пакетної обробки та моніторингу, який інтегровано у фронтенд платформи на Next.js. Виконано функціональне тестування на трьох репрезентативних класах атак (SQL-ін’єкція, програма-вимагач, перебір паролів) та порівняльне оцінювання якості відповідей за метрикою перекриття ключових слів, яке засвідчило суттєве підвищення якості після донавчання порівняно з базовою моделлю. Запропонований підхід відтворюваний на споживчому графічному обладнанні та може бути масштабований для промислового застосування. У статті наведено постановку проблеми, аналіз сучасних промислових та академічних рішень, теоретичні основи методу, методику формування датасету й донавчання, отримані результати та напрями подальших досліджень.

Ключові слова: штучний інтелект; великі мовні моделі; донавчання; QLoRA; кібербезпека; аналіз інцидентів; MITRE ATT&CK; центр управління безпекою; FastAPI; Llama 3.2.

ВСТУП

Постановка проблеми. Стрімке зростання кількості та складності кібератак у 2023-2025 роках поставило перед організаціями нові виклики у сфері інформаційної безпеки. За даними галузевих звітів, середня вартість витоку даних у світі досягла рекордних 4,88 млн доларів США, а середній час виявлення та локалізації інциденту перевищує 250 днів [11]. Центри управління безпекою (Security Operations Centers, SOC) щоденно опрацьовують від десятків до сотень тисяч сповіщень, переважну

більшість яких становлять хибнопозитивні спрацювання, що породжує ефект «втоми від алертів» і підвищує ризик пропустити реальну атаку. Глобальний дефіцит кваліфікованих фахівців із кібербезпеки оцінюється в мільйони осіб [12], а багатоетапні цілеспрямовані атаки (APT) залишають розрізнені сліди в різних системах, що вкрай ускладнює їх ручну кореляцію. Великі мовні моделі (Large Language Models, LLM) пропонують принципово новий підхід: вони здатні опрацьовувати неструктурований текстовий опис інциденту (лог, алерт, звіт) і генерувати структурований аналіз із рекомендаціями, наближений до роботи досвідченого аналітика.

Дослідження виконано в межах проєкту Cybersecurity Dataspace (CSDS) – децентралізованої платформи для обміну даними про кіберінциденти між організаціями на основі блокчейну. Розроблений у роботі AI-модуль розширює функціональність платформи можливістю автоматизованого аналізу поданих інцидентів безпосередньо на локальному обладнанні, без передавання чутливих даних до зовнішніх хмарних сервісів.

Аналіз останніх досліджень і публікацій. Провідні технологічні компанії вже інтегрували LLM-асистентів у власні продукти для кібербезпеки. Microsoft Security Copilot побудовано на основі GPT-4 та архітектури Retrieval-Augmented Generation (RAG) із доступом до власного threat intelligence; рандомізоване дослідження засвідчило підвищення точності аналізу та швидкості роботи аналітиків [3, 4, 10]. IBM QRadar застосовує машинне навчання для поведінкового аналізу та навчання моделей безпосередньо на даних клієнта, що підвищує конфіденційність. Google SecLM обрав підхід повного донавчання на security-специфічному корпусі, а CrowdStrike Charlotte AI спеціалізується на телеметрії кінцевих точок. Спільним обмеженням цих рішень є їхня закритість та залежність від хмари.

Академічна спільнота активно досліджує застосування LLM у SOC. Компанія Sophos опублікувала набір бенчмарків для оцінювання LLM у задачах аналізу, узагальнення та визначення серйозності інцидентів і показала, що domain-specific донавчання суттєво покращує якість [9]. Бенчмарк SECURE підтвердив перевагу комерційних моделей, водночас вказавши, що донавчання відкритих моделей здатне нівелювати цей розрив. Основою практичної реалізації стала робота з QLoRA [2], автори якої довели, що 4-бітна квантизація у поєднанні з адаптерами LoRA [1] забезпечує якість, порівнянну з повним донавчанням, знижуючи вимоги до відеопам’яті у 3-4 рази. Питанням безпеки самих моделей та інструкційного донавчання присвячено праці [13-15, 20], а трансформерну архітектуру як підґрунтя сучасних LLM описано в [16].

Невирішеною частиною загальної проблеми залишається відсутність відкритих, відтворюваних та конфіденційних рішень, придатних для організацій, які не можуть використовувати комерційні хмарні LLM через регуляторні чи бюджетні обмеження. Саме на усунення цієї прогалини спрямовано дане дослідження.

Метою статті є розробка та інтеграція AI-модуля для автоматизованого аналізу інцидентів кібербезпеки на платформі CSDS на основі донавчання великої мовної моделі методом QLoRA. Для досягнення мети розв’язано такі завдання: проаналізовано сучасні промислові й академічні рішення; обґрунтовано архітектурний підхід; сформовано та верифіковано навчальний датасет із відкритих джерел; проведено донавчання базової LLM; розроблено REST API-сервіс для інтеграції моделі з платформою та виконано функціональне тестування й оцінювання якості відповідей.

ТЕОРЕТИЧНІ ОСНОВИ ДОСЛІДЖЕННЯ

При проєктуванні AI-асистента розглянуто чотири основні підходи. Zero-shot/few-shot промптинг готової LLM через API не потребує витрат на тренування, але створює залежність від зовнішнього сервісу та ризики конфіденційності. Архітектура RAG поєднує LLM із векторною базою знань, забезпечуючи актуальність та посилання на джерела, проте потребує складної інфраструктури. Повне донавчання (full fine-tuning) дає максимальну якість, але вимагає значних обчислювальних ресурсів і несе ризик «катастрофічного забування». Параметрично-ефективне донавчання (PEFT), зокрема QLoRA, додає невелику кількість навчальних параметрів-адаптерів за замороженої квантизованої базової моделі.

Метод LoRA ґрунтується на спостереженні, що матриця оновлень ваг під час донавчання має низький внутрішній ранг, тому її подають як добуток двох матриць малого рангу $\Delta W = BA$, де $r << \min(d, k)$, що різко скорочує кількість навчальних параметрів [1]. QLoRA розширює LoRA 4-бітною квантизацією NF4 (NormalFloat), оптимізованою для нормально розподілених ваг, подвійною квантизацією констант та paged-оптимізаторами для уникнення переповнення пам’яті [2]. Унаслідок цього модель розміром 3 млрд параметрів, яка у форматі BF16 потребувала б близько 6 ГБ відеопам’яті, у режимі 4-бітного QLoRA займає близько 2,1 ГБ, що уможливлює донавчання на графічному процесорі з 8 ГБ відеопам’яті.

Таблиця 1

Порівняльний аналіз підходів до побудови AI-асистентів

Критерій	Zero-shot	RAG	Full FT	QLoRA
Вимоги до GPU	Не потрібен	GPU для inference	8×A100 80 ГБ	GPU з 8 ГБ
Час підготовки	Години	1-2 дні	Тижні	1-3 години
Конфіденційність	Дані у хмарі	Можливо локально	Повністю локально	Повністю локально
Вартість	За запит (API)	Середня	Дуже висока	Низька
Кастомізація	Лише промпт	База знань	Повна	Часткова (адаптер)
Придатність для MVP	Так	Частково	Ні	Так (обрано)

За результатами аналізу обрано метод QLoRA. Вибір зумовлено технічною реалізованістю на наявному обладнанні (GPU з 8 ГБ відеопам'яті), академічною обґрунтованістю методу, автономністю й конфіденційністю локального розгортання, а також можливістю проводити кілька незалежних раундів донавчання на розширюваних датасетах за однакових гіперпараметрів.

МЕТОДИКА ДОСЛІДЖЕННЯ

Формування навчального датасету. Для донавчання обрано стратегію диверсифікованих джерел. Сформовано датасет обсягом 1875 прикладів у форматі JSONL (instruction-tuning, Alpaca-style) із п'яти відкритих джерел: MITRE ATT&CK Enterprise Matrix [5] – 500 прикладів; NVD CVE Database [6] – 200; синтетичні логи інцидентів – 200; відкриті Q&A-датасети кібербезпеки з платформи Hugging Face – 975; правила виявлення Sigma/Elastic [17, 18] – 6. Під час формування дотримано принципів релевантності, різноманітності типів атак, структурованості відповідей (секції Threat Type, Severity, Attack Vector, Recommended Actions), достовірності (посилання на техніки MITRE ATT&CK та ідентифікатори CVE) і відсутності дублікатів.

Датасет пройшов чотири етапи обробки: дедуплікацію за точним збігом полів (видалено 87 дублікатів) та за семантичною схожістю з порогом косинусної подібності 0,85 (видалено 36 near-duplicates); ручну фільтрацію за якістю на вибірці 10% прикладів (видалено 23 некоректні приклади); верифікацію тегів MITRE ATT&CK за актуальною версією бази; токен-аналіз із обрізанням прикладів довжиною понад 1024 токени до максимально допустимого значення.

Вибір базової моделі та конфігурація донавчання. Як базову модель обрано Llama 3.2 3B Instruct [7] – оптимальний компроміс між якістю та вимогами до ресурсів, з уже виконаним інструкційним вирівнюванням та ліцензією, що дозволяє некомерційне дослідницьке використання. Сімейство Llama, починаючи з відкритих базових та діалогових моделей Llama 2 [8], послідовно розвивалося в напрямі вищої якості міркування та доступнішого ліцензування, що робить його зручною основою для domain-specific донавчання. донавчання проведено у два незалежні раунди (v1 та v2) на одній і тій самій базовій моделі з ідентичними гіперпараметрами; мета другого раунду – оцінити вплив більшого та різноманітнішого датасету за однакових умов навчання.

Таблиця 2

Конфігурація QLoRA-донавчання

Гіперпараметр	Значення	Обґрунтування
LoRA rank (r)	16	Баланс ємності та відеопам'яті
LoRA alpha (α)	32	Співвідношення α/r = 2
LoRA dropout	0,05	Запобігання перенавчанню
Target modules	q, k, v, o, gate, up, down	Усі шари attention та MLP
Кількість епох	3	Однакові умови для порівняння
Learning rate/scheduler	$2 \cdot 10^{-4}$ / cosine	Стандарт для QLoRA
Batch size/grad. accum.	2/4	Ефективний batch = 8
Max seq. length	1024	Уміщує 95% прикладів
Квантизація/optimizer	NF4 4-bit + DQ/paged_adamw_8bit	Мінімізація відеопам'яті
Trainable params	≈24,3 млн (0,75%)	r = 16, ті самі модулі

Вибір усіх семи типів проєкційних шарів (query, key, value, output у блоках attention та gate, up, down у MLP) є агресивнішим за стандартний LoRA (лише q та v) і підвищує ємність адаптера за незначного зростання кількості параметрів. Тренування виконано на графічному процесорі з 8 ГБ відеопам'яті (CUDA 12.8, PyTorch 2.x) із бібліотеками transformers, trl, peft та bitsandbytes [19].

Архітектура AI-модуля та інтеграція. Модуль реалізовано за принципом мікросервісної архітектури з трьох шарів: шар моделі (завантажена базова модель із накладеним LoRA-адаптером та 4-бітною квантизацією), шар обробки (токенізація, форматування у chat-template, керування параметрами генерації та постобробка) і шар API (FastAPI-застосунок із валідацією через Pydantic та CORS-middleware). Сервіс надає REST-ендпоінти для аналізу інцидентів, багатокрокового діалогу, пакетної обробки та моніторингу. Інтерфейс AI-асистента інтегровано у фронтенд платформи CSDS на Next.js через типізований TypeScript-клієнт.

Методика тестування. Тестування проведено на трьох рівнях: модульне тестування ендпоінтів, функціональне тестування сценаріїв аналізу та інтеграційне тестування взаємодії фронтенду з сервісом. Функціональне тестування зосереджено на трьох репрезентативних класах атак, що охоплюють різні вектори: вебатаки (SQL-ін'єкція), шкідливе програмне забезпечення (програма-вимагач) і мережеві атаки (перебір паролів). Якість відповідей оцінювали за коректністю ідентифікації типу загрози, правильністю визначення серйозності, релевантністю рекомендацій та наявністю посилань на техніки MITRE ATT&CK, а також за інтегральною метрикою перекриття ключових слів (keyword overlap) відносно еталонних відповідей.

Декларація про використання інструментів ШІ. Допоміжні інструменти штучного інтелекту використано виключно для технічного редагування, структурування та мовностилістичного оформлення рукопису згідно з шаблоном видання. Усі наукові результати, експериментальні дані, їх аналіз та висновки отримані й сформульовані авторами самостійно; автори несуть повну відповідальність за зміст статті.

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Результати донавчання. Перший раунд (v1) виконано на датасеті з 1423 прикладів; протягом 534 кроків (3 епохи) функція втрат на навчанні знизилася з 11,09 до 2,63, точність на рівні токенів досягла 83,0%, час тренування – 71 хв за пікового споживання 2,97 ГБ відеопам'яті. Другий раунд (v2) – незалежний запуск на розширеному датасеті з 1875 прикладів за ідентичних гіперпараметрів: 702 кроки, функція втрат 10,80 → 2,94, точність 80,2%. Обидва адаптери (r = 16, float16) мають розмір 92,8 МБ.

Таблиця 3

Порівняння результатів раундів донавчання v1 та v2

Метрика	Версія v1	Версія v2
Розмір датасету	1423 приклади	1875 прикладів
Початкова функція втрат	11,09	10,80
Фінальна функція втрат	2,63	2,94
Точність на рівні токенів	83,0%	80,2%
Кількість кроків	534	702
Час тренування	71 хв	≈95 хв
Розмір адаптера	92,8 МБ	92,8 МБ

Версія v2 продемонструвала дещо гірші метрики на навчальних даних попри більший обсяг датасету. Це пояснюється вищою гетерогенністю розширеного набору та більшою питомою вагою загальних Q&A-прикладів відносно вузькоспеціалізованих, і свідчить про пріоритетність якості та тематичної концентрації прикладів над їх кількістю для донавчання невеликих моделей.

Функціональне тестування. На сценарії SQL-ін'єкції (WAF-лог із payload «OR 1=1--» та сигнатурою інструмента sqlmap) модель коректно визначила тип загрози як Web Attack – SQL Injection (T1190, тактика Initial Access) і рівень Critical, виявила автоматизоване сканування та чинник підвищеного ризику (вихідний вузол TOR), сформувавши пріоритезовані рекомендації. На сценарії програми-вимагача (EDR-алерт із видаленням тіньових копій через vssadmin, масовим перейменуванням файлів та C2-з'єднанням) модель ідентифікувала техніки T1490, T1486, T1059.001 і T1571 та правильно інтерпретувала передвимагацькі дії. На сценарії перебору паролів (SSH-логи з 1247 невдалими спробами

та подальшою успішною автентифікацією) модель виявила критичний момент компрометації й коректно пріоритезувала ізоляцію сесії перед розслідуванням. У всіх трьох випадках досягнуто коректної ідентифікації типу загрози та рівня серйозності з посиланнями на техніки MITRE ATT&CK.

Оцінювання якості відповідей. Для об'єктивного оцінювання проведено порівняльне тестування базової моделі без адаптера та дообучених версій на єдиній тестовій вибірці за метрикою перекриття ключових слів. донавчання забезпечило підвищення середнього показника якості приблизно на 82% порівняно з базовою моделлю (0,393 проти 0,216), а частка задовільних відповідей зросла з 10% до 66%. Показовим є аналіз CVE-вразливості: базова модель лише переказує надану інформацію (0,286), тоді як дообучена відтворює структуровану відповідь із точним ідентифікатором CVE, вектором CVSS та рівнем критичності (0,744). Це підтверджує, що донавчання навчило модель дотримуватися очікуваного формату відповіді, а не лише відповідати на запит. Нижчі за 0,5 абсолютні значення метрики є типовою характеристикою генеративних моделей, які перефразовують відповіді замість дослівного відтворення.

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

У роботі вирішено задачу розробки та інтеграції AI-асистента для автоматизованого аналізу інцидентів кібербезпеки на платформі CSDS. Проаналізовано сучасні промислові й академічні рішення та встановлено, що ключовим обмеженням для їх відтворення є відсутність публічних датасетів реальних корпоративних логів. Обґрунтовано вибір методу QLoRA, що уможливлює донавчання на споживчому графічному процесорі з 8 ГБ відеопам'яті за 1-3 години з повністю локальним і конфіденційним розгортанням. Сформовано та верифіковано навчальний датасет обсягом 1875 прикладів із п'яти відкритих джерел. Проведено два незалежні раунди донавчання моделі Llama 3.2 3B Instruct; функціональне тестування на трьох репрезентативних сценаріях підтвердило коректність ідентифікації типів загроз і рівнів серйозності з посиланнями на техніки MITRE ATT&CK, а порівняльне оцінювання засвідчило підвищення якості відповідей приблизно на 82% порівняно з базовою моделлю.

Перспективними напрямами подальших досліджень є: розширення датасету реальними SIEM-логами у форматах Windows Event Log та syslog; упровадження RAG-розширення для актуального threat intelligence; оптимізація inference через конвертацію у формат GGUF та розгортання за допомогою vLLM; а також оцінювання моделі на незалежному зовнішньому тестовому наборі для остаточного вибору між версіями адаптерів.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., & Chen, W. (2021). *LoRA: Low-rank adaptation of large language models* (arXiv:2106.09685). arXiv. https://arxiv.org/abs/2106.09685
2. Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems, 36*.
3. Freitas, S., Callahan, A., Bhatt, A., & Bono, J. (2024). *AI-driven guided response for security operation centers with Microsoft Copilot for Security* (arXiv:2407.09017). arXiv. https://arxiv.org/abs/2407.09017
4. Bono, J., & Xu, A. (2024). *Randomized controlled trials for Security Copilot for IT administrators* (arXiv:2411.01067). arXiv. https://arxiv.org/abs/2411.01067
5. MITRE Corporation. (2025). *ATT&CK Enterprise Matrix*. https://attack.mitre.org/matrices/enterprise
6. National Institute of Standards and Technology. (2025). *National Vulnerability Database (NVD): CVE API 2.0*. https://nvd.nist.gov/developers/vulnerabilities
7. Meta AI. (2024). *Llama 3.2: Multilingual large language models*. Hugging Face. https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct
8. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., et al. (2023). *Llama 2: Open foundation and fine-tuned chat models* (arXiv:2307.09288). arXiv. https://arxiv.org/abs/2307.09288
9. Sophos. (2024). *LLM benchmarks for cybersecurity: Evaluation of incident investigation, summarisation and severity evaluation* [Technical report].
10. Microsoft. (2024). *Generative AI and security operations center productivity: Evidence from live operations* [Microsoft Research report].
11. IBM Security, & Ponemon Institute. (2024). *Cost of a data breach report 2024*.
12. ISC2. (2024). *Cybersecurity workforce study 2024*.
13. National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). https://doi.org/10.6028/NIST.AI.100-1

14. Perez, F., & Ribeiro, I. (2022). *Ignore previous prompt: Attack techniques for language models* (arXiv:2211.09527). arXiv. https://arxiv.org/abs/2211.09527
15. Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q. V. (2022). Finetuned language models are zero-shot learners. In *International Conference on Learning Representations (ICLR)*.
16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*.
17. SigmaHQ. (2025). *Sigma detection rules repository*. GitHub. https://github.com/SigmaHQ/sigma
18. Elastic. (2025). *Detection rules repository*. GitHub. https://github.com/elastic/detection-rules
19. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., et al. (2020). HuggingFace's Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38-45). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6
20. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems, 35*.

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 2 (34), 2026
ISSN 2663 – 4023

Daniil Kutsiuk
Student of the Bachelor's programme "Information Systems Vulnerability Analysis"
National University of Kyiv-Mohyla Academy, Kyiv, Ukraine
ORCID:0009-0008-1018-3767
d.kutsiuk@ukma.edu.ua

Kyrylo Gorokhovskyi
Senior Lecturer, Department of Multimedia Systems, Faculty of Informatics
National University of Kyiv-Mohyla Academy, Kyiv, Ukraine
ORCID:0000-0002-6398-5367
fintech.lab@ukma.edu.ua

DEVELOPMENT OF AN AI-ASSISTANT MODULE BASED ON QLoRA FINE-TUNING OF A LARGE LANGUAGE MODEL FOR AUTOMATED CYBERSECURITY INCIDENT ANALYSIS

Abstract. The article addresses the development and integration of an artificial-intelligence module for automated cybersecurity incident analysis within the decentralised Cybersecurity Dataspace (CSDS) platform for sharing cyber-incident data between organisations. The relevance of the study stems from systemic problems of modern security operations centres (SOCs): the excessive volume of security events, the global shortage of qualified specialists, the effect of alert fatigue, and the growing complexity of multi-stage attacks. The aim of the work is to build a specialised large language model (LLM) capable of classifying threats, assessing their severity, and providing structured response recommendations in line with the MITRE ATT&CK and NIST SP 800-61 methodologies. The paper provides a comparative analysis of four approaches to building AI assistants (zero-shot prompting, RAG, full fine-tuning, and parameter-efficient fine-tuning) and substantiates the choice of QLoRA as optimal for local deployment given confidentiality requirements and limited computational resources. A training dataset of 1,875 examples was assembled from five open sources (MITRE ATT&CK, NVD CVE, synthetic incident logs, open cybersecurity Q&A datasets, and Sigma/Elastic detection rules); it underwent deduplication, quality filtering, and verification of MITRE techniques. Two independent rounds of QLoRA fine-tuning of the Llama 3.2 3B Instruct base model were carried out with identical hyperparameters, and their results were described in terms of loss and token-level accuracy. A FastAPI microservice with REST endpoints for incident analysis, dialogue, batch processing, and monitoring was developed and integrated into the platform's Next.js front end. Functional testing on three representative attack classes (SQL injection, ransomware, and brute force) and a comparative evaluation of answer quality using a keyword-overlap metric demonstrated a substantial quality improvement after fine-tuning compared with the base model. The proposed approach is reproducible on consumer-grade GPU hardware and can be scaled for industrial use. The article presents the problem statement, an analysis of current industrial and academic solutions, the theoretical basis of the method, the dataset-construction and fine-tuning methodology, the obtained results, and directions for further research.

Keywords: artificial intelligence; large language models; fine-tuning; QLoRA; cybersecurity; incident analysis; MITRE ATT&CK; security operations center; FastAPI; Llama 3.2.

REFERENCES (TRANSLATED AND TRANSLITERATED)

1. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., & Chen, W. (2021). *LoRA: Low-rank adaptation of large language models* (arXiv:2106.09685). arXiv. https://arxiv.org/abs/2106.09685
2. Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems, 36*.
3. Freitas, S., Callahan, A., Bhatt, A., & Bono, J. (2024). *AI-driven guided response for security operation centers with Microsoft Copilot for Security* (arXiv:2407.09017). arXiv. https://arxiv.org/abs/2407.09017
4. Bono, J., & Xu, A. (2024). *Randomized controlled trials for Security Copilot for IT administrators* (arXiv:2411.01067). arXiv. https://arxiv.org/abs/2411.01067
5. MITRE Corporation. (2025). *ATT&CK Enterprise Matrix*. https://attack.mitre.org/matrices/enterprise
6. National Institute of Standards and Technology. (2025). *National Vulnerability Database (NVD): CVE API 2.0*. https://nvd.nist.gov/developers/vulnerabilities

7. Meta AI. (2024). *Llama 3.2: Multilingual large language models*. Hugging Face. https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct
8. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., et al. (2023). *Llama 2: Open foundation and fine-tuned chat models* (arXiv:2307.09288). arXiv. https://arxiv.org/abs/2307.09288
9. Sophos. (2024). *LLM benchmarks for cybersecurity: Evaluation of incident investigation, summarisation and severity evaluation* [Technical report].
10. Microsoft. (2024). *Generative AI and security operations center productivity: Evidence from live operations* [Microsoft Research report].
11. IBM Security, & Ponemon Institute. (2024). *Cost of a data breach report 2024*.
12. ISC2. (2024). *Cybersecurity workforce study 2024*.
13. National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). https://doi.org/10.6028/NIST.AI.100-1
14. Perez, F., & Ribeiro, I. (2022). *Ignore previous prompt: Attack techniques for language models* (arXiv:2211.09527). arXiv. https://arxiv.org/abs/2211.09527
15. Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q. V. (2022). Finetuned language models are zero-shot learners. In *International Conference on Learning Representations (ICLR)*.
16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*.
17. SigmaHQ. (2025). *Sigma detection rules repository*. GitHub. https://github.com/SigmaHQ/sigma
18. Elastic. (2025). *Detection rules repository*. GitHub. https://github.com/elastic/detection-rules
19. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., et al. (2020). HuggingFace's Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38-45). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6
20. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems, 35*.

Отримано редакцією журналу / Received: 26.01.26
Прорецензовано / Revised: 06.02.26
Схвалено до друку / Accepted: 24.09.26