НАЦІОНАЛЬНА АКАДЕМІЯ НАУК УКРАЇНИ

ІНСТИТУТ ПРОБЛЕМ МОДЕЛЮВАННЯ В ЕНЕРГЕТИЦІ ІМ. Г.Є. ПУХОВА

ІНСТИТУТ ПРОБЛЕМ РЕЄСТРАЦІЇ ІНФОРМАЦІЇ

ЗБІРНИК МАТЕРІАЛІВ

НАУКОВО-ПРАКТИЧНОЇ КОНФЕРЕНЦІЇ «ШТУЧНИЙ ІНТЕЛЕКТ І БЕЗПЕКА»

19-21 листопада 2024 р.

Київ–2024

УДК 004(8+056+413.4)
ББК 32.813
Ш-94

Рекомендовано до друку Вченою радою Інституту проблем моделювання в енергетиці ім. Г.Є. Пухова НАН України (протокол № 12 від 28 листопада 2024 р.)

Ш-94 **Штучний інтелект і безпека,** науково-практична конференція Інституту проблем моделювання в енергетиці ім. Г.Є. Пухова Національної академії наук України, Інституту проблем реєстрації інформації Національної академії наук України : матеріали, 19-21 листопада 2024 р. Київ : ІПМЕ ім. Г.Є.Пухова НАН України, ІПРІ НАН України, 2024. 115 с.

SH-94 **Artificial intelligence and security**, scientific-practical conference of the G.E. Pukhov Institute for Modeling in Energy Engineering National Academy of Sciences of Ukraine, Institute for Information Recording of the National Academy of Sciences of Ukraine : materials, November 19-21, 2024. Kyiv: PIMEE NAS of Ukraine, IPRI NAS of Ukraine, 2024. 115 p.

Д.В. Ланде, В.І. Полуциганова, С.А. Смирнов

МЕТОДОЛОГІЯ РОЮ ВІРТУАЛЬНИХ ЕКСПЕРТІВ ДЛЯ ОЦІНКИ ВЗАЄМОЗВ'ЯЗКУ ЗАГРОЗ ТА УРАЗЛИВОСТЕЙ ОБ'ЄКТУ КРИТИЧНОЇ ІНФРАСТРУКТУРИ

У сучасних дослідженнях у сфері оцінки ризиків для об'єктів критичної інфраструктури важливим є методологічний підхід до оцінки взаємозв'язку між загрозами та уразливостями [1], [2]. Традиційно для цих цілей використовуються методи, засновані на експертних оцінках, однак такі підходи часто стикаються з проблемами суб'єктивності, обмеженого числа експертів і складності в інтеграції різноманітних знань.

Одним із підходів, який дозволяє ефективно вирішувати ці проблеми, є методологія «рою віртуальних експертів» [3], [4]. Термін «рій» у цьому контексті означає сукупність численних віртуальних агентів (експертів), які одночасно взаємодіють із системою штучного інтелекту (зокрема, великими мовними моделями - LLM), з метою отримання максимально точних і збалансованих оцінок. Кожен промпт, сформульований як запит до LLM, виступає як окремий "віртуальний експерт", що вносить свій внесок у загальний результат.

Метою роботи є створення та обґрунтування методології рою віртуальних експертів для оцінки взаємозв'язку між загрозами та уразливостями об'єктів критичної інфраструктури з використанням великих мовних моделей (LLM) та їх математичного моделювання для підвищення точності і надійності оцінок ризиків у сфері кібербезпеки.

Особливістю використання рою віртуальних експертів є те, що цей процес контролюється людиною, яка координує запити до системи, формує промпти і інтегрує відповіді, отримані від різних агентів. Це дозволяє зберігати певний рівень контролю над процесом, уникати системних помилок, що можуть виникати через відсутність нагляду.

Принцип роботи рою віртуальних експертів полягає у наступному:

- Кожен віртуальний експерт отримує запит у вигляді певного промпта, що містить інформацію про загрози, уразливості, або інші параметри системи.

- Після того, як кілька експертів надають свої оцінки, ці оцінки агрегуються. Агрегація може бути здійснена за допомогою зваженого середнього, або більш складних методів, таких як методи зліплення оцінок на основі ймовірності.

- Хоча кожен експерт є автономним, роль людини в цьому процесі полягає в тому, щоб здійснювати контроль за формулюванням запитів і корекцією відповідей, а також оцінювати консистентність отриманих результатів. Людина може коригувати промпти або вводити додаткові уточнення.

Використання рою віртуальних експертів для оцінки загроз та уразливостей має такі переваги, як урахування множинності точок зору, забезпечення точності оцінок, гнучкість і адаптивність, з ниження впливу людського фактора. Автоматизація процесу оцінки загроз та уразливостей знижує ймовірність помилок, які можуть бути пов'язані з людським фактором. Водночас, роль людини залишається важливою для контролю за коректністю результатів.

Загрози та уразливості є основними компонентами в аналізі безпеки критичних інфраструктур. Загроза визначається як можливість здійснення негативного впливу на систему, викликаного зловмисними діями або природними факторами. Уразливість — це слабкість системи, яка може бути використана для реалізації загрози. Взаємодія між цими двома поняттями є основою для прогнозування потенційних ризиків та побудови адекватних заходів безпеки.

Для коректного аналізу взаємозв'язків між загрозами та уразливостями пропонується використовувати матриці інцидентності, де кожен елемент матриці позначає наявність зв'язку між конкретною загрозою та уразливістю.

Нехай існує набір загроз $U = \{u_1, u_2, ..., u_m\}$ та набір уразливостей $B = \{b_1, b_2, ..., b_n\}$, де m — кількість загроз, а n — кількість уразливостей. Тепер можемо побудувати матрицю інцидентності M розміру $m \times n$, елементи якої m_{ij} показують наявність взаємозв'язку між загрозою u_i та уразливістю b_j:

$$M = \begin{pmatrix} m_{11}\ m_{12}\ \ldots m_{1n} \\ m_{21}\ m_{22}\ \ldots m_{2n} \\ . \\ . \\ . \\ m_{m1}\ m_{m2}\ \ldots m_{mn} \end{pmatrix},$$

де:

- $m_{ij} = 1$, якщо загроза u_i може бути реалізована через уразливість b_j;
- $m_{ij} = 0$, якщо зв'язку між загрозою та уразливістю немає.

Для зниження похибок та підвищення точності необхідно використовувати методи агрегації відповідей, отриманих від різних LLM. Одним із підходів є використання **методів зваженого середнього**, де кожній

відповіді від конкретної моделі надається вага в залежності від її надійності або точності. Ваги можуть бути визначені на основі досвіду моделей, їх специфікацій або попереднього тестування.

Щоб ефективно агрегувати відповіді від різних LLM у концепції «рою віртуальних експертів», метод зваженого середнього можна організувати із урахуванням:

- Кількості токенів у у кожній LLM. Цей параметр можна використати для надання вищої ваги таким моделям.
- Новіші релізи LLM можуть володіти більш актуальними знаннями, враховувати сучасні технології і методи, а також мати покращену архітектуру.
- Релевантності відповідей на тестових запитах, попередні результати тестування або експертні оцінки.

Після визначення вагових значень для кожної моделі (наприклад, $w_1, w_2, \ldots, w_n$), де w_i – вага відповідної LLM, можна розрахувати середню відповідь з урахуванням внеску кожної моделі. Якщо кожна модель видає оцінку або текстову відповідь a_i, тоді зважена середня відповідь A обчислюється як:

$$A = \frac{\sum_{i=1}^{n} w_i \cdot a_i}{\sum_{i=1}^{n} w_i}.$$

Таким чином, відповіді моделі з вищою вагою більше впливатимуть на кінцевий результат, підвищуючи точність і зменшуючи похибки.

Рій віртуальних експертів генерує числові оцінки на основі зазначених факторів. У таблиці інцидентності кожна клітинка відображає оцінку ймовірності зв'язку між конкретною загрозою та уразливістю. За допомогою методу "середнього" можна агрегувати оцінки рою для кожної клітинки таблиці. Середня оцінка для кожної пари загроза-уразливість визначається як середнє значення по всіх оцінках віртуальних експертів:

$$\hat{m}_{ij} = \frac{1}{K} \sum_{k=1}^{K} m_{ij}^{k}.$$

Така середня оцінка дає загальне уявлення про ймовірність наявності зв'язку між загрозою і уразливістю на основі оцінок рою.

Оскільки оцінки для кожної пари загроза-уразливість генеруються різними віртуальними експертами, важливо оцінити точність середніх оцінок. Можна ввести критерій точності для кожної пари загроза-уразливість, що дозволяє з'ясувати, наскільки точними є середні оцінки.

Для цього введемо функцію точності, що визначає різницю між середньою оцінкою та фактичною оцінкою, яку дають людські експерти або система автоматичної перевірки:

$$\delta_{ij} = \left| \hat{m}_{ij} - \hat{m}_{ij}^{true} \right|,$$

де $\hat{m}_{ij}^{true}$ — оцінка, отримана від людських експертів або іншої перевіреної системи.

Для оцінки ефективності рою віртуальних експертів можна використовувати критерії, такі як середня квадратична помилка (MSE), що дозволяє порівняти середні оцінки з реальними даними:

$$MSE = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} \left(\hat{m}_{ij} - \hat{m}_{ij}^{true} \right)^2,$$

де N і M — кількість загроз та уразливостей відповідно.

Для практичного прикладу використовуємо звіт компанії Edgescan [5], де наведено статистика найбільш поширених уразливостей. За результатами проведеного експертного аналізу сформовано перелік загроз $\{u_i\}$ і уразливостей $\{b_j\}$, які позначають уразливості. Останні взаємопов’язано із потенційно вразливими компонентами, на які саме і можуть бути спрямовані кібератаки.

Для зазначеного об’єкта захисту сформовано наступний перелік елементів, які характеризують уразливості:

b_1 – уразливість вхідних драйверів вхідної інформації;

b_2 – уразливість драйверів інструментів обробки інформації;

b_3 – уразливість драйверів мікросхем BIOS; …

Визначений перелік загроз від кібератак наведено нижче:

u_1 – загроза завантаження шкідливого (вірусного) програмного забезпечення, використовуючи особливості альтернативної операційної системи з розширеними повноваженнями;

u_2 – загроза несанкціоноване копіювання інформації;

u_3 – загроза неавторизованої модифікації інформації; …

Для отримання оцінок від віртуальних експертів, які було реалізовано на базі застосування сервісів ChatGPT (https://chatgpt.com/), Groq (https://groq.com/, модель Llama-3), DeepSeec (https://www.deepseek.com/), застосовувався промпт:

Промпт: *Маємо: деякого об'єкта критичної інфраструктури сформовано наступний перелік елементів, які характеризують уразливості:*
b_1 – уразливість вхідних драйверів вхідної інформації;
b_2 – уразливість драйверів інструментів обробки інформації;
…
Визначений перелік загроз від кібератак на цей обєкт наведено нижче:
u_1 – загроза завантаження шкідливого (вірусного) програмного забезпечення, використовуючи особливості альтернативної операційної системи з розширеними повноваженнями;
u_2 – загроза несанкціоноване копіювання інформації;
…
Створить таблицю інцидентності, де рядки –(u), стовпці – (b).

Матриця інцидентності, яку було розглянуто у прикладі (Рис. 1), виглядає узгодженою та логічною, але для повної впевненості в її коректності бажано перевірити кожен зв'язок на відповідність конкретним технічним сценаріям та умовам об'єкта.

	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}
u_1	**1**										**1**		**1**
u_2				**1**					**1**	**1**	**1**		
u_3		**1**		**1**			**1**	**1**	**1**	**1**	**1**		
u_4				**1**		**1**	**1**						
u_5	**1**	**1**	**1**	**1**					**1**				
u_6					**1**	**1**	**1**	**1**	**1**				
u_7					**1**	**1**	**1**	**1**	**1**	**1**			
u_8												**1**	
u_9												**1**	
u_{10}										**1**			**1**
u_{11}											**1**		
u_{12}					**1**	**1**							
u_{13}					**1**	**1**		**1**					
u_{14}					**1**	**1**	**1**	**1**					

Рисунок 1 - Агрегована таблиця інцидентності

У наведеному прикладі людиною-експертом було підтверджено її логічність, при чому враховувалось декілька критеріїв для оцінки її коректності, а саме логічність зв'язків, повнота матриці, надмірність зв'язків, відповідність реальним практикам безпеки.

Висновки

Наукова новизна цієї роботи полягає в інтеграції рою віртуальних експертів із використанням LLM для автоматизованої, але контрольованої оцінки взаємозв'язків між загрозами і уразливостями в об'єктах критичної інфраструктури, а також у математичному моделюванні цього процесу. Цей підхід дозволяє підвищити точність і надійність оцінок, а також оптимізувати процес прийняття рішень у сфері кібербезпеки.

Метод рою віртуальних експертів є ефективним інструментом для оцінки таблиці інцидентності загроз та уразливостей у системах критичної інфраструктури. Використання таких методів дозволяє агрегувати численні оцінки і надавати середні значення, що відображають ймовірність виникнення конкретної загрози через певну уразливість.

Математичне моделювання процесу оцінки, а також перевірка коректності середніх оцінок через порівняння з оцінками людських експертів, дозволяє підтвердити ефективність і точність застосовуваного методу. Середні оцінки, отримані від рою віртуальних експертів, можуть використовуватись як основа для подальшої оцінки ризиків і планування заходів захисту для критичної інфраструктури.

Цей підхід також є адаптивним, оскільки можна змінювати параметри моделі та адаптувати її під конкретні потреби кожної інфраструктури. Тому метод рою віртуальних експертів може стати важливим інструментом у сфері кібербезпеки.

1. Aslan, Ömer, Semih Serkant Aktuğ, Merve Ozkan-Okay, Abdullah Asim Yilmaz, and Erdal Akin. "A comprehensive review of cyber security vulnerabilities, threats, attacks, and solutions." Electronics 12, no. 6 (2023): 1333. DOI: 10.3390/electronics12061333.

2. Ghelani, Diptiben, Tan Kian Hua, and Surendra Kumar Reddy Koduru. "Cyber security threats, vulnerabilities, and security solutions models in banking." Authorea Preprints (2022). DOI: 10.22541/au.166385206.63311335/v1.

3. Lande D, Strashnoy L. GPT Semantic Networking: A Dream of the Semantic Web - The Time is Now. - Kyiv: Engineering, 2023. - 168 p. ISBN 978-966-2344-94-3.

4. Lande D, Strashnoy L. Swarm of Virtual Experts in the Implementation of Semantic Networking. ResearchGate Preprint, 2024. Access mode: https://doi.org/10.13140/RG.2.2.16686.11845.

5. Vulnerability Statistics Report 2023. Edgescan. URL: https://www.edgescan.com/intel-hub/stats-report/ (date of access: 22.06.2023).

ЗМІСТ