Міжнародний конкурс студентських робіт зі штучного інтелекту 2026

II тур
Перший етап

Р Е Ц Е Н З І Я

на наукову роботу

назва Розробка та оптимізація моделей обробки природної мови для україномовного AI асистента локального користувача в реальному часі

шифр КАПРІКА

представлену на Міжнародний конкурс студентських робіт зі штучного інтелекту

№ з/п	Характеристики та критерії оцінки рукопису наукової роботи	Рейтингова оцінка. Максимальна кількість балів (за 120-бальною шкалою)	Бали
1	Актуальність проблеми	10	9,8
2	Новизна та оригінальність ідей	15	8,0
3	Використані методи дослідження	15	10,5
4	Теоретичні наукові результати	10	4,5
5	Практична направленість результатів	20	16,0
6	Рівень використання наукової літератури та інших джерел інформації	5	3,8
7	Ступінь самостійності роботи	10	8,0
8	Якість оформлення	5	4,3
9	Наукові публікації	10	5,3
10	Недоліки роботи (пояснення зниження максимальних балів у пунктах 1-9):		

Наукова новизна полягає переважно в адаптації та поєднанні відомих підходів до аналізу SIEM-сповіщень. Не наведено абляційного порівняння аналізу окремих сповіщень і часових груп, а також звичайних та стандартизованих промптів. Тому внесок кожного із запропонованих удосконалень у підсумкову якість не встановлено.

Правила часового групування описано, однак вибір інтервалів 5, 15, 30 і 60 хвилин недостатньо обґрунтовано експериментально. Не досліджено чутливості результатів до зміни цих параметрів, ризику об’єднання непов’язаних подій та розділення одного інциденту між кількома епізодами.

Еталонні мітки встановлено аналітиком, проте не наведено незалежної перевірки іншими експертами та оцінки узгодженості розмітки. Клас critical визначає потребу в перевірці або розслідуванні, тому результати безпосередньо характеризують первинне сортування сповіщень, а не підтверджене виявлення завершених кіберінцидентів.

Тестова вибірка сформована зі спеціально зміненим співвідношенням рівнів сповіщень та типів подій. Такий підхід корисний для порівняння моделей, але показники Precision і скорочення потоку сповіщень не можна безпосередньо переносити на повний робочий потік. Потрібне додаткове оцінювання на послідовній часовій вибірці зі збереженням природного розподілу подій.

Висновок про найкращий баланс GPT-OSS 120B потребує визначення вартості різних типів помилок. Модель пропустила 30 із 61 критичного епізоду. Її збалансована точність практично дорівнює показнику GPT-OSS 20B – 74,95% проти 74,90%, але профілі помилок суттєво відрізняються. Перевагу слід визначати відповідно до допустимого ризику пропуску та навантаження на аналітика.

Не наведено порівняння з базовою процедурою сортування за рівнем правил Wazuh або іншим простим класифікатором. Через це не встановлено, який додатковий ефект забезпечує застосування LLM порівняно з уже доступною інформацією SIEM. Теоретичні результати переважно обмежуються описом процедури та відомих метрик.

10	Необхідно однозначно описати врахування технічних помилок, відповідей insufficient_evidence та направлень на ручну перевірку. Матриці помилок містять по 500 класифікованих епізодів для кожної моделі, тоді як окремо наведено технічні відмови та неоднозначні рішення. Потрібно пояснити їх перетворення на підсумкові класи, кількість повторних звернень і вплив відбору успішних відповідей на метрики. Перевірка стійкості на 11 епізодах є корисною, але недостатньою для узагальнення ефективності голосування більшості. Не наведено критеріїв відбору цієї підвибірки, довірчих інтервалів та повторної перевірки на всіх 500 епізодах. Порівняння часових характеристик також залежить від різних провайдерів, обмежень частоти звернень і повторних спроб, тому не характеризує лише швидкодію моделей. Вимоги щодо доказовості та ігнорування сторонніх інструкцій у журналах доцільні, проте їх фактичну ефективність не перевірено окремими тестами. Не наведено кількісного оцінювання правильності пояснень, посилань на докази та зіставлення з MITRE ATT&CK. Огляд джерел доцільно посилити прямим порівнянням із найближчими дослідженнями LLM-сортування SIEM-сповіщень. Оформлення загалом забезпечує розуміння роботи, але окремі знімки інтерфейсу містять дрібний текст; наявні повтори та редакційні неточності. Цінність роботи – використання реальних журналів подій, продуманий набір метрик і порівняння кількох мовних моделей в однакових умовах. Відзначимо такі недоліки: методика складається з відомих кроків, і з роботи не видно, яке саме нововведення дає найбільший ефект (окремої перевірки групування та уніфікованих запитів немає); залишається відкритим питання, чи перевершує мовна модель звичайну фільтрацію за критичністю правил Wazuh; такий простий еталон для порівняння варто було додати; еталонну розмітку виконано однією особою, а вибірку збагачено подіями високого рівня, тому отримані показники не характеризують роботу на звичайному робочому потоці; модель, визнана найкращою, залишає непоміченою приблизно половину критичних епізодів, що для засобу кібербезпеки суттєво (доцільно обговорити, яка частка пропусків є прийнятною); для організацій критичної інфраструктури надсилання журналів у сторонні хмарні сервіси навряд чи допустиме, тому варто дослідити локальні моделі. Варто чіткіше обґрунтувати, для яких типів кіберінцидентів і організаційних умов запропонована методика є найбільш ефективною. Новизна має переважно прикладний характер. Доцільно чіткіше відмежувати запропоновану методику від наявних підходів до кореляції подій і застосування LLM у кібербезпеці. Варто детальніше описати формування набору епізодів, розподіл класів інцидентів, критерії розмітки «еталонних» рішень та процедуру запобігання упередженості при оцінюванні відповідей моделей. Теоретичне узагальнення є досить обмеженим. Доцільно формальніше описати умови застосування методу, критерії групування подій за часом, правила інтерпретації відповідей LLM і межі їхньої надійності. Наукова новизна має переважно описово-методичний характер. Відсутні методи донавчання, не досліджено вплив розміру контекстного вікна або методів пошуку, а відбір епізодів для розмітки містить частку суб'єктивізму. Теоретичний апарат у роботі мінімальний. Передача внутрішніх журналів безпеки критичної інфраструктури до сторонніх закордонних хмарних API створює критичні ризики витоку чутливих даних організації. Значну частину складають посилання на онлайн-документацію провайдерів API. Розроблені програмні модулі мають високу практичну цінність для CSIRT-підрозділів і служб інформаційної безпеки. Для впровадження важливо додатково оцінити стійкість методики до шуму, неповних журналів, зловмисних маніпуляцій із журналом подій та різних форматів SIEM-систем. Доцільно ширше залучити новітні публікації щодо LLM у кібербезпеці, виявлення аномалій, кореляції подій, prompt engineering та оцінювання надійності ШІ-систем. Варто чіткіше розмежувати власний внесок автора й використані готові LLM, API, бібліотеки, SIEM-рішення та інструменти обробки журналів. Доцільно покращити наочність подання результатів порівняння моделей: додати зведені таблиці, графіки метрик і приклади коректних та помилкових рішень LLM. Варто детальніше описати архітектуру, інтеграцію з конкретними SIEM-системами, залежності, умови розгортання та відтворюваність експериментів. Тестування на реальних журналах CSIRT-ED є досить вагомим практичним результатом. Водночас доцільно навести повніші кількісні результати: точність, повноту, F1-міру, частоту помилкових рішень, стабільність формату відповідей, час обробки та порівняльні таблиці для всіх досліджених моделей.		
10.1	Наявність розробленого програмного забезпечення, у т.ч. веб-сервісу, мобільного застосунку	10	5,3
10.2	Рівень отриманих практичних результатів наукової роботи	10	5,0
Сума балів			80,3

Загальний висновок не рекомендується для захисту на ПНПК

(рекомендується, не рекомендується для захисту на науково-практичній конференції)

Рецензент *члени Конкурсної комісії напряму «Обробка природної мови»*

«06» листопада 2026 року