

РЕЦЕНЗІЯ
Тецького Артема Григоровича
на дисертаційну роботу Неретіна Олексія Сергійовича
на тему «Методи та засоби аналізу кібербезпеки і захисту великих мовних моделей від
генерації забороненого контенту на локальних і хмарних серверах»,
подану на здобуття ступеня доктора філософії
у галузі знань 12 – Інформаційні технології
за спеціальністю 125 – Кібербезпека

1. Актуальність теми дисертації

Стрімкий прогрес великих мовних моделей та їхнє впровадження в освіту, науку й державне управління роблять питання їхньої захищеності надзвичайно гострим. Оскільки ці моделі стають інструментами прийняття рішень у критичних системах, виникає нагальна потреба у розробці методів достовірного оцінювання та гарантування їхньої кіберстійкості.

Актуальність роботи зумовлена браком наукових напрацювань в питаннях практичної безпеки великих мовних моделей. Дослідження Неретіна О. С. пропонує системний підхід до підвищення рівня кіберзахисту мовних моделей через підвищення повноти оцінювання кібербезпеки та рівня їх захищеності. Це дозволяє ефективно протидіяти генерації забороненого контенту великими мовними моделями, які розташовані у локальних та хмарних середовищах, що відповідає пріоритетним запитам у захисті цієї технології.

2. Оцінка обґрунтованості наукових результатів дисертації, їх достовірності та новизни

Наукова новизна результатів дисертаційного дослідження полягає в наступному:

1. Вперше запропоновано модель кібербезпеки великих мовних моделей, яка, на відміну від відомих, надає теоретико-множинне представлення загроз, вразливостей та кібератак на модельному та системному рівнях, що надає змогу здійснювати подальший аналіз ризиків порушення, оцінювати рівень захищеності та визначати контрзаходи.

2. Удосконалено метод аналізу критичності вразливостей великих мовних моделей шляхом вибору джерел даних з експлойтами, їх колекціонування та симулювання атакуючих моделей для статистичної оцінки ймовірності та успішності атак, а також її комбінування з рівнем тяжкості наслідків для ризик-орієнтованого визначення критичності, що забезпечує підвищення повноти та достовірності оцінювання кібербезпеки.

3. Дістав подальшого розвитку IMECA (Intrusion Modes and Effects Criticality Analysis) метод оцінювання та забезпечення кібербезпеки великих мовних моделей шляхом аналізу наслідків атак на вразливості та вибору контрзаходів за частковим та узагальненим показниками, що дозволяє гарантувати прийнятний ризик порушення кібербезпеки з урахуванням ресурсних обмежень.

Наукова достовірність результатів базується на використанні перевіреного математичного апарату: теорії множин, теорії ймовірностей та математичного моделювання. Обґрунтованість підходів підтверджена застосуванням методу IMECA-

аналізу та вибору контрзаходів, а також успішною апробацією розроблених алгоритмів під час експериментального оцінювання кібербезпеки сучасних мовних моделей.

Зазначені результати є новими, обґрунтованими та підтвердженими експериментально.

Наукові дослідження були виконані здобувачем на кафедрі кібербезпеки та інтелектуальних інформаційних технологій (503) Національного аерокосмічного університету «Харківський авіаційний інститут» в рамках НДР: «Методи, програмно-апаратні засоби та технології забезпечення гарантоздатності інтелектуальних систем індустриального інтернету речей» (№ Д/Р 0122U001065, 2022-2023 рр.), «Методи, засоби та технології моделювання, розроблення, розгортання та забезпечення гарантоздатності мобільних інтелектуальних систем для об'єктів критичної інфраструктури» (№ Д/Р 0124U003250, 2024-теперішній час).

Таким чином, в дисертаційному дослідженні поставлене наукове завдання «розроблення методів та засобів аналізу критичності вразливостей, прогнозування кібератак на великі мовні моделі, які розгортаються локально і на хмарних серверах, та обґрунтування вибору і впровадження контрзаходів за визначеними критеріями» виконано в повному обсязі, здобувач продемонстрував високий рівень володіння методологією наукової діяльності.

3. Оцінка змісту дисертації, її завершеність та дотримання принципів академічної доброчесності

За своїм змістом дисертаційна робота здобувача Неретіна О. С. повністю відповідає Стандарту вищої освіти зі спеціальності 125 «Кібербезпека» та напрямкам досліджень відповідно до освітньої програми «Кібербезпека».

Представлене дослідження є цілісною науковою працею, у якій відображено вагомий особистий внесок здобувача у розвиток відповідного наукового напрямку.

Аналіз звіту про подібність підтверджує, що дисертація Неретіна Олексія Сергійовича є оригінальною та самостійною науковою працею. У роботі відсутні ознаки академічної недоброчесності (плагіат, фальсифікація чи фабрикація), а всі запозичені ідеї та матеріали супроводжуються коректними бібліографічними посиланнями.

4. Мова та стиль викладення результатів

Дисертаційна робота написана українською мовою. Текст відповідає вимогам наукового стилю. Використання фахової термінології разом із влучним ілюстративним матеріалом робить виклад складних рішень доступним для розуміння. Робота структурована логічно, а її зміст повністю розкриває авторську концепцію. Кожне теоретичне та прикладне положення має належне наукове обґрунтування.

5. Структура роботи

Дисертація складається з вступу, чотирьох розділів, висновків, списку використаних джерел та додатків. Загальний обсяг дисертації становить 183 сторінки.

У вступі обґрунтовано вибір теми дослідження, сформульовано об'єкт, предмет, мету і завдання дослідження, наведено методи дослідження, а також відображено наукову новизну і практичне значення результатів.

У першому розділі проведено аналіз існуючих методів і засобів оцінювання і забезпечення кібербезпеки великих мовних моделей. Виявлено їх основні недоліки та обмеження. Обґрунтовано необхідність забезпечення захисту технології мовних моделей за рахунок розроблення відповідних моделей, методів та засобів.

У другому розділі на базі математичного апарату теорії множин розроблено модель кібербезпеки систем з мовними моделями та модель кібербезпеки безпосередньо великих мовних моделей. Крім того, удосконалено метод аналізу критичності вразливостей великих мовних моделей, який базується на симулюванні атакуювання мовних моделей та дозволяє пом'якшити обмеження представлених у першому розділі методів.

У третьому розділі було адаптовано ІМЕСА метод оцінювання та забезпечення кібербезпеки для аналізу великих мовних моделей, а також вибрано контрзаходи до їх вразливостей, критерії та алгоритми їх вибору за частковим та узагальненим показниками.

Четвертий розділ присвячено розробці та тестуванню технології оцінювання кібербезпеки великих мовних моделей. Доведено ефективність рішення у протидії генерації забороненого контенту. Практична цінність та новизна роботи підкріплені успішними результатами апробації розроблених методів.

У висновках наведено основні результати дисертаційної роботи, сформульовано практичне значення отриманих результатів, а також визначено напрями подальших досліджень.

Дисертаційна робота оформлена відповідно до вимог наказу МОН України від 12 січня 2017 р. № 40 «Про затвердження вимог до оформлення дисертації».

6. Оприлюднення результатів дисертаційної роботи

Наукові результати дисертації висвітлені у 11 наукових публікаціях здобувача, серед яких: 5 статей у наукових виданнях, включених на дату опублікування до переліку наукових фахових видань України; 4 статті у періодичних наукових виданнях, проіндексованих у базах даних Web of Science Core Collection та/або Scopus, з яких 1 стаття у виданнях, віднесених до першого-третього квартилів (Q1–Q3) відповідно до класифікації SCImago Journal and Country Rank або Journal Citation Reports.

Крім того, результати дисертації були апробовані на 8 наукових фахових конференціях.

Публікації здобувача повністю відповідають темі дисертації та вичерпно розкривають її основні положення. Апробація результатів проведена через представлення доповідей на конференціях різного рівня та статті у фахових виданнях. Роботи присвячені методам аналізу кібербезпеки та захисту мовних моделей від генерації забороненого контенту в локальних і хмарних середовищах, що релевантно змісту дослідження.

Дисертація виконана з дотриманням принципів академічної доброчесності: усі джерела вказані коректно, а запозичений матеріал належним чином процитований, що виключає наявність плагіату. У спільних працях роль здобувача була ключовою – вона охоплює теоретичну розробку методів і алгоритмів, а також самостійне проведення експериментів та інтерпретацію даних.

Публікації здобувача вичерпно розкривають зміст дослідження, цілком відповідають фаховим критеріям та підтверджують високу якість проведеного дослідження.

Отже, основні положення та висновки дисертації знайшли повне відображення у наукових публікаціях здобувача.

7. Недоліки та зауваження до дисертаційної роботи

1. В роботі використано визначення рівня тяжкості згідно до законодавства ЄС, але не розкрито питання стосовно можливості адаптації цього рівня під законодавства інших країн.

2. В роботі не надано аналізу ризиків щодо комбінованих атак на системи LLMs.

3. Автор обрав спрощений варіант вибору контрзаходів оскільки розглядає обмежену множину вразливостей та контрзаходів, а також варіантів їх покриття.

4. Не зважаючи на те, що в темі говориться про локальні та віддалені сервери, більша увага була приділена локальним серверам. Тож експериментального підтвердження результатів для віддалених серверів в роботі не надано.

Вважаю, що зазначені зауваження не є визначальними і не зменшують загальну наукову новизну та практичну цінність отриманих результатів і не впливають на загальну позитивну оцінку дослідження.

8. Висновок про дисертаційну роботу

Вважаю, що дисертаційна робота здобувача ступеня доктора філософії Неретіна Олексія Сергійовича на тему «Методи та засоби аналізу кібербезпеки і захисту великих мовних моделей від генерації забороненого контенту на локальних і хмарних серверах» є завершеною та самостійною науковою працею. Робота виконана з дотриманням норм академічної доброчесності, а її результати забезпечують розв'язання актуальної науково-практичної задачі, що має вагомe значення для розвитку галузі інформаційних технологій. Наукова новизна, актуальність та отримані результати роботи повною мірою задовольняють вимогам чинного законодавства України, що передбачені в п.6–9 «Порядку присудження ступеня доктора філософії та скасування рішення разової спеціалізованої вченої ради закладу вищої освіти, наукової установи про присудження ступеня доктора філософії», затвердженого Постановою Кабінету Міністрів України від 12 січня 2022 р. № 44.

Здобувач Неретін Олексій Сергійович заслуговує на присудження ступеня доктора філософії в галузі знань 12 Інформаційні технології за спеціальністю 125 Кібербезпека.

Рецензент:

кандидат технічних наук,
доцент кафедри кібербезпеки та інтелектуальних
інформаційних технологій
Національного аерокосмічного університету
«Харківський авіаційний інститут»

Артем ТЕЦЬКИЙ