

РЕЦЕНЗІЯ

Брежнєва Євгена Віталійовича
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

Актуальність теми дисертації.

Актуальність теми дисертаційної роботи зумовлена стрімким розвитком технологій штучного інтелекту, зокрема великих мовних моделей (ВММ). Велика кількість параметрів мовних моделей надає їм змогу аналізувати та розуміти людську мову на дуже високому рівні. Ці моделі використовуються у різних сферах діяльності, таких як освіта та наука, судова та медична системи, критична інфраструктура. Мовні моделі виконують завдання з діагностування, оптимізації та автоматизації різних послуг, підвищення точності та глибшого розуміння. Стрімкий розвиток цієї технології у поєднанні з її використанням у чутливих сферах діяльності обумовлює необхідність достовірного оцінювання і гарантованого забезпечення кібербезпеки мовних моделей.

Відсутність науково обґрунтованих і відпрацьованих в практичному сенсі методів оцінювання та забезпечення кібербезпеки великих мовних моделей робить цей напрям актуальним і важливим для підвищення надійності та якості цієї технології в цілому.

Таким чином, дослідження Неретіна О. С., спрямоване на підвищення повноти оцінювання кібербезпеки та рівня захищеності великих мовних моделей від генерації забороненого контенту шляхом аналізу критичності вразливостей, прогнозування кібератак на моделі, які розгортаються локально та на хмарних серверах, є актуальним та відповідає сучасним потребам у захисті цієї технології.

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 атакування 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.

Достовірність отриманих наукових результатів підтверджується застосуванням загальновідомих методів досліджень, зокрема методу математичного моделювання, теорії множин, теорії ймовірностей, методу IMECA-аналізу та вибору контрзаходів за визначеними критеріями, а також експериментальними прикладами оцінювання та

забезпечення кібербезпеки з використанням розроблених методів і алгоритмів для існуючих 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).

Отже, в дисертаційній роботі 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 повністю, здобувач повною мірою оволодів методологією наукової діяльності.

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».

Дисертація є цілісним та завершеним науковим дослідженням, результати якого відображають індивідуальний внесок автора у розвиток відповідного наукового напрямку.

На підставі результатів перевірки на текстові збіги встановлено, що дослідження виконане автором особисто без використання недозволених методів запозичення (фальсифікації, компіляції, фабрикації чи плагіату). Використання сторонніх наукових джерел у роботі є коректним і відповідає вимогам до цитування.

Мова та стиль викладення результатів.

Дисертаційна робота написана українською мовою з дотриманням норм наукового стилю. Автор використовує коректну термінологію, властиву фаховим і науковим дослідженням у відповідній галузі. Завдяки чітким визначенням, поясненням та ілюстраціям запропоновані рішення стають легшими для сприйняття. Робота відзначається цілісністю викладу, повнотою охоплення теми та чітким розкриттям основної наукової ідеї автора. Представлені у дисертаційній роботі теоретичні та прикладні положення викладені послідовно, логічно та належно обґрунтовані.

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 методів і засобів оцінювання і забезпечення кібербезпеки великих мовних моделей. Виявлено основні недоліки та обмеження існуючих рішень у цій сфері і обґрунтовано необхідність розроблення методів та засобів аналізу кібербезпеки і захисту великих мовних моделей від генерації забороненого контенту на локальних і хмарних серверах.

У другому розділі розроблено модель кібербезпеки систем, які використовують великі мовні моделі, яка є базовою для розроблення спеціалізованої моделі кібербезпеки мовних моделей. Розроблені моделі використовують математичний апарат теорії множин.

Наступним кроком, на базі зазначених моделей, удосконалено метод аналізу критичності вразливостей великих мовних моделей, який допомагає подолати або пом'якшити обмеження розглянутих у першому розділі методів. Розроблені моделі та метод дозволяють проводити подальше оцінювання та забезпечення кібербезпеки мовних моделей з ціллю підвищення якості цієї технології.

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

У четвертому розділі розроблено інформаційну технологію оцінювання та забезпечення кібербезпеки великих мовних моделей, представлено та проаналізовано результати експериментальних досліджень. Підтверджено продуктивність програмного засобу у підвищенні безпеки мовних моделей від загрози генерації забороненого контенту. Засвідчено практичну цінність та підтверджено наукову новизну дослідження на базі аналізу результатів впровадження розроблених методів та засобів.

У висновках наведено основні результати дисертаційної роботи, сформульовано практичне значення отриманих результатів, 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».

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.

Результати дисертації були апробовані на 8 наукових фахових конференціях.

Наукові публікації здобувача повною мірою узгоджуються з темою дослідження та розкривають основні результати дисертації. Результати дослідження пройшли належну апробацію, що підтверджується публікаціями у фахових наукових виданнях та виступами на міжнародних і всеукраїнських конференціях. Тематика публікацій охоплює питання аналізу кібербезпеки і захисту великих мовних моделей від генерації забороненого контенту на локальних і хмарних серверах, що повністю узгоджується зі змістом дисертації.

Робота відповідає вимогам академічної доброчесності, цитування оформлені належним чином, усі запозичення підкріплені посиланнями, ознаки академічного плагіату відсутні.

У публікаціях, виконаних у співавторстві, особистий внесок здобувача є визначальним і полягає у розробленні моделей, методів та алгоритмів, а також у проведенні експериментальних досліджень та аналізі отриманих результатів.

Загалом публікації здобувача належним чином репрезентують результати дисертаційної роботи, відповідають фаховим вимогам та підтверджують високу якість проведеного дослідження.

Таким чином, зміст дисертації та її наукова новизна цілком репрезентовані у переліку наукових публікацій здобувача.

Основні зауваження до дисертаційної роботи.

1. Для більш глибокого оцінювання кібербезпеки ВММ могло бути використано математичні методи, методи теорії ігор (розділ 1.3.1, сторінка 39), які є більш точними ніж евристичні методи. Але автор обрав більш простий апарат теоретико-множинного опису

(розділ 1.3.1, сторінка 39) та ІМЕСА орієнтованого аналізу (розділ 1.3.2, сторінка 44), що є менш формалізованими та більш залежними від людського фактору.

2. В роботі не досить повно розкрито особливості генерації забороненого контенту на локальних та хмарних серверах. Не показано специфічні загрози та методи кіберзахисту ВММ можуть бути застосовані на локальних та хмарних серверах.

3. Робота обмежується суто статичними моделями та методами оцінювання безпеки ВММ систем без урахування часових характеристик різних типів кібератак. Вектори атак, вразливості, загрози та ризики постійно змінюються. Без урахування часу показники кібербезпеки ВММ швидко стають неактуальними.

4. В роботі стверджується щодо результати можуть бути використані для системи моделей ВММ. Мова є про те, що багатоагентні системи (MAS) на основі ВММ дозволяють групам інтелектуальних агентів координувати та вирішувати складні завдання колективно у великих масштабах, переходячи від ізольованих моделей до підходів, орієнтованих на співпрацю. Вони можуть співпрацювати з точку зору кібербезпеки. Аспект взаємодії в роботі не врахований в повному обсязі.

Вважаю, що зазначені зауваження мають переважно уточнювальний характер і не зменшують наукову цінність і практичну вагу отриманих результатів та не впливають на позитивну оцінку дисертаційної роботи.

Висновки. 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» виконана на високому фаховому рівні з дотриманням усіх норм академічної доброчесності. Автор представив сукупність обґрунтованих теоретичних і практичних рішень, які мають суттєвий вплив на сучасний стан галузі інформаційних технологій. За рівнем наукової новизни, актуальності та практичної значущості представлена дисертація цілком відповідає нормативним вимогам, встановленим чинним законодавством 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:

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

Євген БРЕЖНЄВ