

## ВІДГУК

офіційного опонента Каштальян Антоніни Сергіївни  
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.** Стрімкий прогрес великих мовних моделей відкриває нові можливості для освіти, медицини та ІТ, проте несе в собі приховані загрози. Здатність цих моделей генерувати людиноподібний текст іноді супроводжується токсичністю, упередженістю або небезпечними порадами. Через масштабне впровадження цієї технології критично важливо зосередитись на оцінці ризиків та зміцненні їхньої кібербезпеки.

Автор зазначає, що на практиці бракує перевірених методів захисту великих мовних моделей. Тому виникає логічна потреба розробити надійні методи та інструменти кібербезпеки, які б відповідали стрімкому поширенню цієї технології.

Отже, створення цілісного апарату для оцінювання вразливостей, прогнозування атак та вибору оптимальних засобів захисту є нагальною науково-практичною задачею, спрямованою на безпеку використання мовних моделей.

**Оцінка обґрунтованості наукових результатів дисертації, їх достовірності та новизни.** Наукова новизна результатів дисертаційного дослідження полягає в такому:

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 ІМЕСА (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 наукових результатів ґрунтується на застосуванні математичного апарату, зокрема теорії множин, теорії ймовірностей та методу ІМЕСА, що в поєднанні з результатами експериментальних досліджень безпеки реальних мовних моделей підтверджує об'єктивність висновків. Практична значущість роботи доведена успішною апробацією на локальних мовних моделях, що відкриває широкі перспективи для

використання результатів у корпоративному сегменті та подальших науково-прикладних розробках.

Зазначені результати дисертаційного дослідження є новими, обґрунтованими та підтвердженими шляхом експериментального атакування мовних моделей.

Отже, в дисертаційній роботі поставлене наукове завдання 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».

Дисертаційна робота є завершеною науковою працею і свідчить про наявність особистого внеску здобувача у розвиток відповідного наукового напрямку.

Розглянувши звіт подібності за результатами перевірки дисертаційної роботи на текстові збіги, можна зробити висновок, що дисертаційна Неретіна Олексія Сергійовича є результатом самостійних досліджень здобувача і не містить елементів фальсифікації, компіляції, фабрикації, плагіату та запозичень. Використані ідеї, результати і тексти інших авторів мають належні посилання на відповідне джерело.

**Мова та стиль викладення результатів.** Дисертаційна робота написана українською мовою та відповідає вимогам наукового стилю. Структура роботи відзначається логічною послідовністю: від глибокого аналізу предметної області до практичної апробації розроблених моделей і методів. Це забезпечує цілісність та обґрунтованість отриманих результатів. Матеріал викладено професійною мовою з чітким дотриманням термінології кібербезпеки, водночас складні теоретичні положення залишаються доступними для розуміння.

**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 дисертаційного дослідження, сформульовано об'єкт, предмет, мету і завдання дослідження, наведено методи дослідження, а також відображено наукову новизну і практичне значення результатів.

Перший розділ присвячено критичному огляду засобів гарантування кібербезпеки великих мовних моделей та виявленню їхніх обмежень. Аргументовано доцільність розроблення методів та засобів аналізу їх кібербезпеки і захисту від генерації забороненого контенту.

У другому розділі розроблено загальну модель кібербезпеки систем з великими мовними моделями та спеціалізовану модель кібербезпеки мовних моделей, які базуються на математичному апараті теорії множин. Крім того, удосконалено метод аналізу критичності вразливостей великих мовних моделей, який пом'якшує обмеження розглянутих у першому розділі методів аналізу. Розроблені моделі та метод відкривають шлях до проведення оцінювання та забезпечення кібербезпеки мовних моделей, ціллю якого є підвищення якості технології мовних моделей в цілому.

Другий розділ представляє теоретико-множинні моделі кібербезпеки для систем із великими мовними моделями та безпосередньо цих мовних моделей. Удосконалено метод

аналізу їх вразливостей, що дозволяє пом'якшити обмеження відомих методів. Впровадження цих моделей та методу забезпечує комплексний підхід до оцінювання безпеки мовних моделей та надає можливість подальшого забезпечення захищеності цієї технології.

У третьому розділі удосконалено метод ІМЕСА з метою його адаптації до специфіки великих мовних моделей. На основі проведеного аналізу визначено перелік контрзаходів для усунення вразливостей цих моделей, а також сформульовано критерії та алгоритми їхнього вибору за визначеними показниками.

У четвертому розділі розроблено програмне забезпечення для комплексного оцінювання та забезпечення кіберзахисту великих мовних моделей. Експериментально доведено ефективність створеного інструменту в запобіганні генерації забороненого контенту. Аналіз результатів впровадження підтвердив наукову новизну дослідження та його значущість для практичного застосування у сфері безпеки мовних моделей.

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

Також результати дисертації були апробовані на 8 наукових фахових конференціях.

Опубліковані праці автора цілком розкривають зміст дисертаційного дослідження та аргументують його ключові положення. Усі матеріали підготовлені з дотриманням принципів академічної етики та правил цитування.

У спільних публікаціях особистий внесок здобувача є визначальним: ним сформульовано мету, розроблено методологічну базу та проведено фінальний аналіз результатів.

Таким чином, наукові результати описані в дисертаційній роботі повністю висвітлені у наукових публікаціях здобувача.

#### **Недоліки та зауваження.**

1. В роботі не надано рекомендацій щодо використання LLMs моделей для підсилення захисту комп'ютерних систем і мереж.

2. З тексту дослідження залишається незрозумілим, який саме тип мовних моделей (відкриті або закриті) брався за основу. Потребує пояснення, чи аналізувався вплив архітектурного типу моделей на результати експериментів, зокрема в контексті генерації забороненого контенту.

3. В роботі розглянута вразливість статистично ймовірнісної генерації відповіді але, при цьому, моделі розроблені для множини вразливостей систем та моделей LLMs. Є певна суперечність в описі вразливостей, які враховуються.

4. У роботі зазначено, що методи оцінювання та забезпечення кібербезпеки мовних моделей можуть бути адаптовані до різних прикладних сфер, зокрема для безпілотних літальних апаратів. Разом з тим, цей напрям визначено у загальному вигляді (теоретично), без детального опису процесу цієї адаптації.

5. Експериментальна частина дослідження виконана виключно з використанням локально розгорнутих мовних моделей. У роботі не вистачає порівняльного аналізу використання розробленої методології оцінювання та забезпечення кібербезпеки моделей, розташованих на хмарних серверах. Відсутність інформації про специфіку експериментування у хмарах не дозволяє оцінити особливості функціонування запропонованих методів у контексті мережових затримок, масштабування моделей та специфічних загроз безпеки хмарних моделей.

Вважаю, що висловлені зауваження не є концептуальними, не зменшують загальної наукової новизни та практичної значимості результатів і не впливають на позитивну оцінку дисертаційної роботи.

**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» виконана на високому науковому рівні, не порушує принципів академічної доброчесності та є закінченим науковим дослідженням, сукупність теоретичних та практичних результатів якого розв'язує наукове завдання, що має істотне значення для галузі інформаційних технологій. Дисертаційна робота за актуальністю, практичною цінністю та науковою новизною повністю відповідає 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.

#### **Офіційний опонент:**

професор кафедри комп'ютерної інженерії та  
інформаційних систем  
Хмельницького національного університету,  
доктор технічних наук, доцент

Антоніна КАШТАЛЬЯН