

ВІДГУК

офіційного опонента Яцківа Василя Васильовича
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
з галузі знань 12 Інформаційні технології
за спеціальністю 125 Кібербезпека

Актуальність теми дисертації.

Стрімкий розвиток великих мовних моделей зумовлює зростання інтересу до них у багатьох сферах людської діяльності. З кожним наступним оновленням зростають їх можливості в розумінні людської мови та генерації тексту, подібного до людського. Ця технологія активно використовується в освіті, медицині та індустрії розробки програмного забезпечення, що сприяє значному прогресу в цих галузях. Але, незважаючи на значний прогрес, досягнутий завдяки використанню цих моделей, вони можуть поводитися несподівано, не так, як було задумано їх розробниками. Моделі можуть образити, бути упередженими та надавати шкідливі поради у професійних сферах. З огляду на широке використання моделей у різних галузях та потенційні наслідки від цього, важливо оцінити можливі ризики та забезпечити прийнятний рівень кібербезпеки для цієї технології.

Автор звертає увагу на відсутність обґрунтованих і відпрацьованих в практичному сенсі методів оцінювання та забезпечення кібербезпеки великих мовних моделей, а також зазначає, що прагнення подолати протиріччя між активним зростанням застосувань мовних моделей та відсутністю концептуальних моделей і досконалих методів оцінювання та забезпечення їх кібербезпеки є об'єктивним.

Таким чином, розроблення та дослідження комплексу моделей та методів аналізу критичності вразливостей цих моделей, прогнозування кібератак та обґрунтування вибору і впровадження контрзаходів за визначеними критеріями є актуальною задачею, яка націлена на захист цієї технології.

Зв'язок з науковими програмами, темами.

Дисертаційна робота виконана у Національному аерокосмічному університеті «Харківський авіаційний інститут» відповідно до державних програм та планів НДР, зокрема: НДР «Методи, програмно-апаратні засоби та технології забезпечення гарантоздатності інтелектуальних систем індустріального інтернету речей» (номер державної реєстрації 0122U001065,

2022-2023 рр.); НДР «Методи, засоби та технології моделювання, розроблення, розгортання та забезпечення гарантоздатності мобільних інтелектуальних систем для об'єктів критичної інфраструктури» (номер державної реєстрації 0124U003250, 2024-теперішній час).

Наукова новизна.

Основні наукові здобутки, отримані автором, включають:

- 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;

- 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;

- 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.

Оцінка обґрунтованості наукових результатів дисертації, їх достовірності та новизни.

Достовірність отриманих наукових результатів підтверджується як використанням методу математичного моделювання, теорії множин, теорії ймовірностей та методу ІМЕСА-аналізу, так і експериментальними прикладами оцінювання та забезпечення кібербезпеки існуючих великих мовних моделей.

Практична цінність роботи підтверджується апробацією запропонованих підходів на реальних локальних мовних моделях, що підтверджує можливість використання її результатів у майбутніх прикладних дослідженнях, науково-дослідних проєктах та у корпоративному секторі.

Отримані результати дисертації характеризуються науковою новизною, є ґрунтовними та верифіковані за допомогою експериментальних атак на мовні моделі.

Таким чином, дисертаційне дослідження вирішує поставлене наукове завдання в повному обсязі, автор 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».

Представлене дослідження є цілісною науковою працею, результати якої демонструють значущість особистого внеску автора для відповідної галузі знань.

Аналіз звіту про подібність підтверджує, що дисертація Неретіна Олексія Сергійовича є самостійною науковою працею. У роботі не виявлено ознак плагіату, фальсифікації чи інших порушень академічної доброчесності, а всі запозичені ідеї та тексти супроводжуються коректними посиланнями.

Мова та стиль викладення результатів.

Дисертаційна робота написана українською мовою та відповідає вимогам наукового стилю. Матеріал викладено логічно та послідовно: від аналізу предметної області до розроблення моделей, методів і практичного підтвердження їх продуктивності, що робить отримані результати цілісними та зрозумілими. Матеріал подано доступно, точно та аргументовано. Складні концепції супроводжуються зрозумілими поясненнями, а спеціалізована термінологія зі сфери кібербезпеки вжита професійно та доречно.

Загальна характеристика дисертаційної 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.

У першому розділі проведено аналіз існуючих методів і засобів оцінювання і забезпечення кібербезпеки великих мовних моделей, а також виявлено їх основні недоліки та обмеження. Обґрунтовано необхідність забезпечення захисту мовних моделей, що зумовлює розроблення методів та засобів аналізу їх кібербезпеки і захисту від генерації забороненого контенту на локальних і хмарних серверах.

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.

Третій розділ присвячено подальшому розвитку ІМЕСА методу оцінювання та забезпечення кібербезпеки. Проведено адаптацію цього методу для аналізу стану кібербезпеки технології великих мовних моделей. Вибрано контрзаходи до вразливостей цих моделей, критерії та алгоритми їх вибору за визначеними показниками.

Четвертий розділ присвячено розробленню програмного засобу для оцінювання та забезпечення кібербезпеки великих мовних моделей, який дозволить проводити комплексне оцінювання стану захищеності технології мовних моделей, а також забезпечувати прийнятний рівень їх кібербезпеки. Проведено експериментальні дослідження, на базі яких підтверджено продуктивність програмного засобу у підвищенні захищеності мовних моделей від загрози генерації забороненого контенту. За результатами аналізу впровадження розроблених методів та засобів засвідчено практичну цінність дослідження та підтверджено його наукову новизну.

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.

Матеріали дисертаційної роботи були представлені та обговорені на 8 наукових фахових конференціях.

Наукові публікації автора повністю відповідають темі дисертації та підкріплюють її результати. Роботи виконані самостійно, з коректним цитуванням та дотриманням усіх вимог академічної доброчесності.

У роботах, написаних у співавторстві, здобувач відіграв провідну роль, забезпечивши постановку дослідницьких завдань, розробку моделей та методів, а також узагальнення отриманих висновків.

Отже, зміст дисертації та її основні результати пройшли належну апробацію та повною мірою відображені в опублікованих працях автора.

Недоліки та зауваження до дисертаційної роботи

1. В дисертаційній роботі розглядаються питання захисту інформації на локальних та віддалених серверах. Зрозуміло, що там використовуються усталені методи захисту, однак, на мій погляд, доцільно було б надати рекомендації стосовно можливості використання криптографічних методів для захисту LLMs систем з урахуванням їх особливостей.

2. В роботі в явному вигляді не надається визначення та розрахунки для показника стійкості LLMs моделей до кібератак.

3. Автор не використовує моделі, які можуть оцінити зміни показників кібербезпеки в часі (з урахуванням часових рядів для кібератак).

Зазначені зауваження мають уточнювальний характер і не заперечують наукову цінність та практичну вагу роботи, а отже, не змінюють її загальну позитивну оцінку.

Висновок про дисертаційну роботу

Вважаю, що представлена 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 ґрунтовним дослідженням, теоретичні та практичні результати якого розв'язують наукове завдання, що має істотне значення для розвитку галузі інформаційних технологій. Дотримання принципів академічної доброчесності, актуальність, практична цінність та наукова новизна роботи повністю відповідають 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 № 44 від 12 січня 2022 р., а її автор, **Неретін Олексій Сергійович** заслуговує на присудження ступеня доктора філософії в галузі знань 12 Інформаційні технології за спеціальністю 125 Кібербезпека.

Офіційний опонент:

завідувач кафедри кібербезпеки

Західноукраїнського національного університету,

доктор технічних наук, професор

Василь ЯЦКІВ