ВИСНОВОК

про наукову новизну, теоретичне та практичне значення результатів дисертації *Неретіна Олексія Сергійовича* на тему «Методи та засоби аналізу кібербезпеки і захисту великих мовних моделей від генерації забороненого контенту на локальних і хмарних серверах», представлену на здобуття ступеня доктора філософії в галузі знань 12 Інформаційні технології за спеціальністю 125 Кібербезпека

На засіданні кафедри кібербезпеки та інтелектуальних інформаційних технологій за участі:

Харченка Вячеслава Сергійовича, чл.-кор. НАН України, д.т.н., професора, завідувача кафедри кібербезпеки та інтелектуальних інформаційних технологій «ХАІ»;
Фесенка Германа Вікторовича, д.т.н., професора, професора кафедри кібербезпеки та інтелектуальних інформаційних технологій «ХАІ»;
Брежнєва Євгена Віталійовича, д.т.н., професора кафедри кібербезпеки та інтелектуальних інформаційних технологій «ХАІ»;
Пєвнєва Володимира Яковлевича, д.т.н., доцента, професора кафедри кібербезпеки та інтелектуальних інформаційних технологій «ХАІ»;
Клюшнікова Ігоря Миколайовича, д.т.н., ст. наук. співр., доцента кафедри кібербезпеки та інтелектуальних інформаційних технологій «ХАІ»;
Перепелицина Артема Євгеновича, к.т.н., доцента, доцента кафедри кібербезпеки та інтелектуальних інформаційних технологій «ХАІ»;
Колісник Марини Олександрівни, к.т.н., доцента, доцента кафедри кібербезпеки та інтелектуальних інформаційних технологій «ХАІ»;
Піскачова Олександра Івановича, к.т.н., ст. наук. співр., доцента кафедри кібербезпеки та інтелектуальних інформаційних технологій «ХАІ»;
Бабешка Євгена Васильовича, к.т.н., доцента, доцента кафедри кібербезпеки та інтелектуальних інформаційних технологій «ХАІ»;
Дужого В'ячеслава Ігоровича, к.т.н., доцента, доцента кафедри кібербезпеки та інтелектуальних інформаційних технологій «ХАІ»;
Тецького Артема Григоровича, к.т.н., доцента кафедри кібербезпеки та інтелектуальних інформаційних технологій «ХАІ»;
Куланова Віталія Олександровича, к.т.н., доцента кафедри кібербезпеки та інтелектуальних інформаційних технологій «ХАІ»;
Землянка Георгія Андрійовича, д.ф., доцента кафедри кібербезпеки та інтелектуальних інформаційних технологій «ХАІ»;
Вдовіченка Олександра Олександровича, д.ф., доцента кафедри кібербезпеки та інтелектуальних інформаційних технологій «ХАІ»;
Дужої Вікторії Вікторівни, ст. викладача кафедри кібербезпеки та інтелектуальних інформаційних технологій «ХАІ»;
Холодної Зої Борисівни, ст. викладача кафедри кібербезпеки та інтелектуальних інформаційних технологій «ХАІ»;

Желтухіна Олександра Васильовича, ст. викладача кафедри кібербезпеки та інтелектуальних інформаційних технологій «ХАІ»;
Годунова Олександра Сергійовича, ст. викладача кафедри кібербезпеки та інтелектуальних інформаційних технологій «ХАІ»;
Демури Руслана Івановича, аспіранта кафедри кібербезпеки та інтелектуальних інформаційних технологій «ХАІ»;
Семенця Олександра Юрійовича, аспіранта кафедри кібербезпеки та інтелектуальних інформаційних технологій «ХАІ»;
Остапенка Леоніда Юрійовича, аспіранта кафедри кібербезпеки та інтелектуальних інформаційних технологій «ХАІ»;
Абакумова Артема Ігоровича, аспіранта кафедри кібербезпеки та інтелектуальних інформаційних технологій «ХАІ»;
Канарського Євгенія Олександровича, аспіранта кафедри кібербезпеки та інтелектуальних інформаційних технологій «ХАІ»;
Щеглова Владислава Романовича, аспіранта кафедри кібербезпеки та інтелектуальних інформаційних технологій «ХАІ»;
Скоробогатька Станіслава Віталійовича, аспіранта кафедри кібербезпеки та інтелектуальних інформаційних технологій «ХАІ»;
Юдіна Олеся Вікторовича, аспіранта кафедри кібербезпеки та інтелектуальних інформаційних технологій «ХАІ»;
Косаревського Богдана Валерійовича, аспіранта кафедри кібербезпеки та інтелектуальних інформаційних технологій «ХАІ»;
Стряпуніна Антона Олександровича, аспіранта кафедри кібербезпеки та інтелектуальних інформаційних технологій «ХАІ»;

а також запрошених:
Соколової Євгенії Віталіївни, к.т.н., доцента, доцента кафедри інженерії програмного забезпечення «ХАІ»;
Федоренка Миколи Івановича, к.т.н, асистента кафедри прикладної лінгвістики «ХАІ»;
Заславського Володимира Анатолійовича, д.т.н., професора, професора кафедри математичної інформатики Київського національного університету імені Тараса Шевченка;
Волочія Богдана Юрійовича, д.т.н., професора, професора кафедри програмно-апаратних систем інфокомунікацій Національного університету «Львівська політехніка»;
Каштальян Антоніни Сергіївни, д.т.н, доцента, професора кафедри комп'ютерної інженерії та інформаційних систем Хмельницького національного університету;
Яцківа Василя Васильовича, д.т.н, професора, завідувача кафедри кібербезпеки Західноукраїнського національного університету;
Коваля Василя Сергійовича, к.т.н, доцента кафедри інформаційно-обчислювальних систем та управління Західноукраїнського національного університету;

Бикового Павла Євгеновича, к.т.н, доцента кафедри інформаційно-обчислювальних систем та управління Західноукраїнського національного університету,

відбулася публічна презентація дисертаційної роботи *Неретіна Олексія Сергійовича* на тему «**Методи та засоби аналізу кібербезпеки і захисту великих мовних моделей від генерації забороненого контенту на локальних і хмарних серверах**».

На підставі обговорення змісту презентації дисертаційної роботи ухвалено такий висновок про наукову новизну, теоретичне та практичне значення результатів дисертації (результати голосування – одноголосно).

1. Актуальність теми дослідження

Актуальність теми дослідження зумовлена активним використанням великих мовних моделей чутливими сферами суспільства, такими як освітня та наукова сфери, судова та медична системи, критична інфраструктура та інші. Ця технологія використовується для виконання завдань діагностування, аналізу даних, оптимізації та автоматизації різних послуг, підвищення точності та глибшого розуміння, що обумовлює необхідність достовірного оцінювання і гарантованого забезпечення кібербезпеки цих моделей. Водночас існуючі підходи не забезпечують повноту оцінювання безпеки та рівня захищеності мовних моделей, що, обумовлює необхідність розроблення та дослідження комплексу моделей та методів систематизації та аналізу критичності вразливостей цих моделей, прогнозування кібератак на великі мовні моделі, які розгортаються локально і на хмарних серверах, та обґрунтування вибору і впровадження контрзаходів за визначеними критеріями.

2. Зв'язок роботи з науковими програмами, планами, темами

Отримані автором результати дисертації виконано на кафедрі кібербезпеки та інтелектуальних інформаційних технологій Національного аерокосмічного університету «ХАІ» в рамках виконання держбюджетних науково-дослідницьких робіт «Методи, програмно-апаратні засоби та технології забезпечення гарантоздатності інтелектуальних систем індустріального 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).

3. Наукова новизна отриманих результатів

У дисертації одержані такі нові наукові результати:

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 статичної 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 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.

4. Теоретичне та практичне значення результатів роботи

Теоретичне значення результатів роботи полягає у розвитку підходів до оцінювання та забезпечення кібербезпеки великих мовних моделей шляхом поєднання моделі їх кібербезпеки, методу аналізу критичності їх вразливостей та ризик орієнтованого методу IMECA, що дозволяє проводити структуроване та формалізоване оцінювання та забезпечення прийнятного ризику порушення їх кібербезпеки за наявності обмежень, а також раціонально розподіляти ресурси для протидії найкритичнішим загрозам у динамічному середовищі мовних моделей.

Практичне значення результатів роботи полягає в тому, що на їх основі були розроблені алгоритми та програмні засоби для проведення IMECA аналізу кібербезпеки великих мовних моделей, структура та елементи інформаційної технології для ризик-орієнтованого оцінювання та вибору контрзаходів для забезпечення безпеки LLMs відповідно до вимог.

Отримані наукові результати можуть бути використані в компаніях-розробниках та провайдерах сервісів великих мовних моделей, аудиторських компаніях та організаціях-розробниках стандартів щодо використання великих мовних моделей, науково-дослідних проєктах та навчальному процесі кафедри 503.

5. Апробація/використання результатів дисертації

Основні результати роботи представлені на конференціях:

1. "Критичні комп'ютерні технології та системи (КриКТехС-2022/6/171)" (м. Харків, Україна, 2022 р.);

2. "Dependable System, Services and Technologies Conference" (Athens, Greece, 2022, 2025);

3. Матеріали II НТК "Інформаційна, функціональна та кібербезпека (СКІФіК-2022)" (м. Харків, Україна, 2022 р.);

4. "Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications" (Dortmund, Germany, 2023);

5. "Digital Interaction and Machine Intelligence" (Warsaw, Poland, 2024);

6. "Global Security Transformation Towards 2040: Transcendents in the Age of AI" (Sofia, Bulgaria, 2025);

7. "Integrated Computer Technologies in Mechanical Engineering" (м. Харків, Україна, 2025 р.).

Результати дисертаційного дослідження впроваджено:

– у навчальному процесі кафедри кібербезпеки та інтелектуальних інформаційних технологій у вигляді лекційного матеріалу та лабораторних занять у навчальній дисципліні «Штучний інтелект і бази знань» (4 години), зокрема, під час розгляду підходів до аналізу вразливостей, ризик-орієнтованого оцінювання кіберзахищеності та вибору контрзаходів для великих мовних моделей і LLM-систем, розроблення програмних засобів їх розгортання та проведення експериментальних досліджень при виконанні кваліфікаційних робіт бакалаврів і магістрів кафедри за спеціальністю «Кібербезпека та захист інформації»;

– при виконанні науково-дослідницьких робіт «Методи, програмно-апаратні засоби та технології забезпечення гарантоздатності інтелектуальних систем індустріального 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);

– при аналізі кібербезпеки великих мовних моделей у компанії ТОВ «ВЕБСПЕЛЧЕКЕР».

6. Дотримання принципів академічної доброчесності

Дисертація О. С. Неретіна є оригінальною роботою, виконана здобувачем самостійно й доброчесно, текст рукопису дисертаційної роботи на містить ознак академічного шахрайства. Роботу передано експерту для проведення науково-технічної експертизи щодо збігів з Internet-джерелами, про що буде надано відповідний звіт.

7. Перелік публікацій за темою дисертації із зазначенням особистого внеску здобувача.

За результатами досліджень опубліковано 11 наукових публікацій, у тому числі:

– 4 статті у наукових фахових виданнях України категорії «Б» за спеціальністю;

– 1 стаття у фаховому виданні України категорії «А», проіндексована в базі Scopus (квартиль Q2);

– 1 стаття опублікована у періодичному виданні Springer (має ISSN та DOI з індексацією у Scopus, квартиль Q4);

– 1 розділ в колективній монографії;

– 2 публікації у просідінгах міжнародних конференцій з індексацією у Scopus;

– 1 публікація у матеріалах міжнародної конференції;

– 1 публікація у матеріалах національної конференції.

Стаття в фаховому виданні України категорії «А», проіндексована в базі Scopus (квартиль Q2):

1. Neretin, O., Kharchenko, V. A model of ensuring LLM cybersecurity. *Radioelectronic and Computer Systems*. 2025. No. 2. P. 201-215. DOI: 10.32620/reks.2025.2.13.

У статті розроблено та проведено аналіз компонентів моделі кібербезпеки LLMs для підвищення точності їх оцінювання та забезпечення необхідного рівня безпеки. Важливість розробки цієї моделі полягає в тому, що вона є базовою для всіх наступних досліджень. Практичне значення аналізу елементів моделі полягає в їх використанні для проведення експериментів з моделювання кібератак на LLMs.

Статті в наукових виданнях України категорії «Б», затверджених як фахові за спеціальністю 125:

1. Неретін, О., Харченко, В. Забезпечення кібербезпеки систем штучного інтелекту: аналіз вразливостей, атак і контрзаходів. *Вісник Національного університету "Львівська політехніка". Інформаційні системи та мережі*. 2022. Т. 12. С. 7–22. DOI: 10.23939/sisn2022.12.007.

У статті проведено огляд стану сучасних підходів до забезпечення кібербезпеки систем штучного інтелекту, класифіковано можливі типи атак і детально розглянуто основні з них, проаналізовано загрози і атаки за рівнем тяжкості і оцінено ризики безпеки з використанням методу IMECA, обґрунтовано напрями подальших досліджень щодо необхідності розроблення методів оцінювання і забезпечення кібербезпеки систем штучного інтелекту.

2. Неретін О., Харченко В. Метод аналізу критичності вразливостей великих мовних моделей. *Вимірювальна та обчислювальна техніка в технологічних процесах*. 2026. № 1. С. 443–450. DOI: 10.31891/2219-9365-2026-85-54.

У статті представлено удосконалений метод аналізу критичності вразливостей LLMs, розгорнутих до використання, визначено основні кроки цього методу, а саме: колекціонування експлойтів до вразливостей моделей, за допомогою яких здійснюється аналіз критичності ризиків; визначення рівня тяжкості наслідків атакування LLMs, що базується на суворості покарання згідно до законодавства Європейського Союзу; проведення симуляції атакування задля визначення статистичної оцінки ймовірності появи та успішності атаки; визначення рівня критичності ризиків як комбінації статистичної оцінки ймовірності та тяжкості наслідків атакування.

3. Neretin, O., Kharchenko, V. IMECA method of risk-based assessment and ensuring cybersecurity of Large Language Models. *Herald of Advanced Information Technology*. 2026. Vol. 9, no. 1. P. 60–70. DOI: 10.15276/hait.09.2026.05.

У статті пропонується метод протидії загрозі генерації забороненого контенту великими мовними моделями шляхом оцінювання ризиків такої

поведінки та забезпечення прийнятного рівня кібербезпеки для LLMs за допомогою методики IMECA. Розроблено набір контрзаходів для підвищення рівня безпеки LLMs та визначено процедури їх вибору на основі критеріїв максимальної продуктивності та найкращого рейтингу за допомогою матриці рейтингу контрзаходів.

4. Neretin, O., Kharchenko, V. Information Technology for Assessing and Ensuring Cybersecurity of Large Language Models. *Security of Infocommunication Systems and Internet of Things*. 2025. Vol. 3, no. 2, paper 02020. P. 1-7. DOI: 10.31861/sisiot2025.2.02020.

У статті представлено програмний інструмент та технологію для оцінювання та забезпечення кібербезпеки LLMs від генерації забороненого контенту. Визначено набір основних даних, необхідних для програмного інструменту, який включає експлойти, підказки для перевірки виводу моделі та контрзаходи для її захисту. Запропоновано процедуру збору, перетворення, зберігання та потенційного розширення та адаптації цих даних до індивідуальних вимог користувачів інструменту. Розроблено функціональну модель технології.

8. Висновок наукового керівника

Виконання індивідуального навчального плану, індивідуального плану наукової роботи, досягнення результатів навчання за відповідною науково-освітньою програмою та написання дисертації Неретіним Олексієм Сергійовичем вважаю успішним. Дисертаційна робота є результатом самостійного дослідження, завершеною науковою працею, яка містить наукову новизну. Вона виконана на високому науковому рівні та відповідає всім установленим вимогам до дисертацій на здобуття наукового ступеня доктора філософії, й може бути рекомендована до захисту, а її автор Неретін Олексій Сергійович – до присудження наукового ступеня доктора філософії за спеціальністю 125 Кібербезпека.

Отже, вважаємо, що дисертаційна робота Неретіна Олексія Сергійовича на тему «Методи та засоби аналізу кібербезпеки і захисту великих мовних моделей від генерації забороненого контенту на локальних і хмарних серверах», представлена на здобуття ступеня доктора філософії, відповідає вимогам Порядку присудження наукового ступеня доктора філософії (Постанова Кабінету Міністрів України від 12 січня 2022 р. №44). Відтак, вона може бути представлена до захисту в разовій спеціалізованій раді для присудження ступеня доктора філософії в галузі знань 12 Інформаційні технології за спеціальністю 125 Кібербезпека.

Головуючий на засіданні
доктор технічних наук, старший науковий співробітник,
доцент кафедри кібербезпеки та
інтелектуальних інформаційних технологій
Національного аерокосмічного університету
«Харківський авіаційний інститут» Ігор КЛЮШНІКОВ

03.04.2026 р.