

## РЕЦЕНЗІЯ

доктора технічних наук, професора

**Федорова Євгена Євгеновича**

на дисертацію *Никоненка Андрія Олександровича*

за темою *«Гібридний метод ідентифікації автоматично згенерованих природномовних текстів»*,

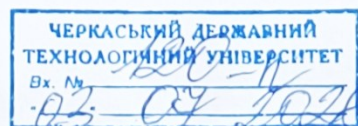
яку подано на здобуття ступеня доктора філософії  
за спеціальністю 122 «Комп'ютерні науки» галузі знань  
12 «Інформаційні технології»

### 1. Актуальність теми дисертаційного дослідження

Актуальність дослідження зумовлена значним зростанням кількості моделей ШІ, які є у вільному доступі, підвищенням їх якості та запровадженням і широкою адаптацією нових інструментів на базі LLM. Покращення інтелектуальних можливостей моделей та розширення списку галузей знань, у яких вони здатні оперувати, веде до ширшого застосування ШІ і, як результат, до збільшення кількості згенерованих даних. Кілька змін технологічних парадигм, що відбулися в галузі, призвели до суттєвих покращень в архітектурі моделей, що дозволило створювати значно більші й потужніші моделі разом зі зниженням вартості їх використання. Доступність застосування ШІ разом із суттєвим підвищенням якості згенерованого контенту дали поштовх зростанню ринку прикладних інструментів на базі ШІ. На сьогодні використання засобів ШІ є де-факто стандартом у різних галузях професійної діяльності та повсякденного життя. Зростання розміру моделей надало ШІ більше можливостей щодо запам'ятовування патернів у людських текстах і дозволило наблизити згенерований контент до написаного людиною.

Широке проникнення генеративного ШІ в людське життя разом із високою якістю створеного контенту призвело до ситуації, коли навіть експерту складно однозначно визначити джерело даних. У такому випадку застосовуються спеціально створені системи-детектори, що дозволяють ідентифікувати сигнал ШІ та маркувати створені ним тексти. Основна потреба в таких технологіях продиктована як захистом якості та достовірності інформації, так і протидією зловживанням ШІ, наприклад, в академічному середовищі, а також захистом від ризиків масштабного розповсюдження дезінформації, питаннями кібербезпеки та соціальної інженерії.

Тому створення гібридного методу ідентифікації автоматично згенерованих природномовних текстів, який здатний демонструвати високу ефективність ідентифікації джерел текстів, протистояти різним типам трансформаційних атак, виявляти стійкість до комерційних систем маскування сліду ШІ та зменшувати ризики хибних звинувачень, є актуальним та своєчасним науково-прикладним завданням.



## 2. Наукова новизна отриманих результатів

Наукова новизна дисертаційної роботи полягає в розробці та обґрунтуванні моделей і гібридного методу ідентифікації автоматично згенерованих природномовних текстів, що суттєво підвищує стійкість детекції до трансформаційних атак та забезпечує прозору інтерпретованість прийнятих рішень в умовах невизначеності.

**Вперше** розроблено гібридний метод ідентифікації автоматично згенерованих текстів, який, завдяки застосуванню каскадного ансамблю з механізмами адаптивного зважування та аналізу міжрівневої узгодженості ознак, дає змогу виявляти дисонанс між лексичним, синтаксичним та семантичним рівнями тексту, забезпечуючи стійкість до комерційних систем маскування і зміщуючи «Точку зламу» детектора до критичних меж втрати семантичної цілісності.

**Вперше** розроблено інтерпретаційну модель обґрунтування доказів детекції, яка, на відміну від обчислювально ресурсомістких ітеративних ХАІ-методів, використовує внутрішні ваги метакласифікатора для прямої генерації глобального та локальних профілів атрибуції та карт доказів, що дає змогу знизити часову складність побудови пояснень до  $O(1)$  та забезпечити прозорість рішень детектора ІІІ.

**Одержали подальший розвиток** методи забезпечення робастності моделей обробки природної мови в умовах агресивного маскування згенерованого тексту та епістемічної невизначеності шляхом інтеграції фреймворку індуктивного конформного передбачення (ICP), що дає змогу ефективно виявляти аномалії, генерувати порожні множини замість хибних висновків та кардинально знижувати рівень хибнопозитивних спрацювань у сценаріях аналізу даних поза навчальним розподілом.

**Удосконалено** концептуальну модель оцінки робастності систем детекції шляхом введення формалізованого поняття «Точки зламу», яке встановлює математичний взаємозв'язок між інтенсивністю структурних змін тексту, збереженням його семантичної цілісності та ймовірністю виявлення, що дає змогу використовувати модель як прогностичний інструмент оцінки чутливості детектора в умовах невизначеності та активного маскування згенерованого контенту.

## 3. Практичне значення результатів дослідження

Практична цінність дисертації підтверджується створенням та розгортанням в інфраструктурі Amazon Web Services програмного комплексу для детекції текстів, згенерованих ІІІ. Важливим аспектом практичної реалізації є швидкодія створеного рішення, незважаючи на використання трансформерів та великих мовних моделей як компонентів системи, середній час аналізу для одного документа становить усього

0.33 секунди, що дозволяє використовувати систему для масової перевірки текстів у реальному часі.

Варті уваги також отримані результати гібридного методу на стаціонарних даних. Так, показник F1 досяг 92% за умови збереження повноти виявлення на рівні 88.77% навіть за умови обмеження хибнопозитивних спрацювань (FPR) на рівні 1%. В умовах активної протидії, на текстах модифікованих комерційними системами маскування, система продемонструвала беззаперечну перевагу над сучасними аналогами. Під час валідації на текстах, оброблених комерційними системами маскування (StealthGPT, AIHumanize, Phrasly), гібридний метод показав повноту виявлення на рівні 77.16% і отримав найнижчий відсоток успішних атак ASR = 22.84%. Крім того, імплементація методу ICP зробила можливим зниження показника помилкових спрацювань для текстів неносіїв мови з 4.81% до 0.98%, що відповідає Rescue Rate = 79.71%.

#### **4. Аналіз структури та змісту дисертації**

Дисертація характеризується чіткою, збалансованою та продуманою структурою. Робота складається зі вступу, чотирьох розділів, загальних висновків, списку використаних джерел (134 найменування) та п'яти додатків. Загальний обсяг становить 248 сторінок із яких 188 - основний текст. Матеріал проілюстровано 45 рисунками та 40 таблицями.

У вступі послідовно та аргументовано розкрито актуальність проблеми, поставлено мету, визначено завдання, об'єкт і предмет дослідження, а також чітко подано наукову новизну та практичну цінність роботи.

Перший розділ присвячено системному аналізу предметної області. Автор розглядає етапи еволюції генеративних моделей ШІ, формалізує концепцію «сліду ШІ» та проводить детальний аналіз традиційних методів детекції, об'єктивно розбираючи їхні теоретичні обмеження та вразливості.

Другий розділ закладає математичну основу роботи. Здобувач будує багатовимірну модель сліду ШІ, концепцію «Точки зламу» та детально описує архітектуру каскадного ансамблю гібридного методу разом з алгоритмом канонізації тексту, механізмом адаптивного зважування і математичним апаратом для моделі пояснювального ШІ.

Третій розділ присвячено експериментальній перевірці теоретичних гіпотез. У ньому описано методологію моделювання трансформаційних атак з використанням моделей DIPPER та T5, а також проведено валідацію гіпотез щодо поведінки сліду ШІ для таких специфічних архітектур, як MoE та Reasoning-моделі.

Четвертий розділ сфокусовано на практичній реалізації. У розділі наведено порівняльний аналіз гібридного методу із SOTA-аналогами, описано архітектуру розгортання системи в AWS та проведено валідацію роботи індуктивного

конформного передбачення на OOD-даних із демонстрацією прикладних сценаріїв роботи інтерпретаційної моделі.

## **5. Дотримання принципів академічної доброчесності**

Дисертаційне дослідження та матеріали наукових публікацій здобувач виконані на високому рівні академічної етики. Ознак порушення принципів академічної доброчесності, фабрикації даних чи некоректних запозичень не виявлено. Робота є самостійною, оригінальною науковою працею.

## **6. Повнота висвітлення результатів дисертації у наукових публікаціях**

Результати дослідження вичерпно представлені у 10 наукових працях. Здобувач має 1 публікацію в іноземному виданні, що індексується наукометричними базами Scopus та DBLP, 3 статті у фахових виданнях України категорії Б, а також 4 публікації у матеріалах міжнародних та всеукраїнських науково-практичних конференцій. Опублікований доробок повністю відповідає вимогам до дисертацій на здобуття ступеня доктора філософії.

## **7. Зауваження щодо змісту та оформлення дисертації**

Позитивно оцінюючи дисертаційну роботу Никоненка А.О., водночас слід звернути увагу на низку дискусійних положень і зауважень:

- 1) У підрозділі 4.2.2 (с. 194-197) наведено розрахунок оцінки обчислювальної складності системи, яка становить  $O(N^2)$ . Оскільки обчислення семантичної подібності речень через косинусну міру з використанням моделі all-MiniLM-L6-v2 та парсинг дерев залежностей у модулі  $M_{syn}$  є ресурсомісткими операціями, то доцільно було б дослідити можливість застосування легших архітектур, наприклад, 1D-CNN як початкового префільтра вхідного потоку документів. Це дозволило б активувати важкі моделі-трансформери лише для текстів із «сірої зони», що дало б можливість суттєво оптимізувати витрати на інференс у хмарному середовищі.
- 2) У підрозділі 2.3.2 (с. 102-105) здобувач пропонує використовувати досить агресивний алгоритм канонізації та нормалізації тексту шляхом використання формату NFKC для захисту від структурних атак. Однак у роботі недостатньо проаналізовано ризики того, що застосування описаного підходу може занадто згладити текстові дані та видалити частину легітимних пунктуаційних чи синтаксичних маркерів. Це, зі свого боку, може створити додатковий статистичний шум та вплинути на точність синтаксичного модуля  $M_{syn}$ , зокрема, під час розрахунку дивергенції Кульбака-Лейблера.
- 3) У підрозділі 4.3.1 (с. 199-202) обґрунтовується використання локальної моделі

Phi-3-mini-4k-instruct у ролі проксі-джерела для розрахунку лексичних імовірностей. Дослідження набуло б більшої переконливості, якби автор детально проаналізував, наскільки архітектурні обмеження LLM такого розміру та типу, зокрема розмір її словника та механізми уваги, впливають на точність підрахунку перплексії під час роботи з даними, що були згенеровані сучасними флагманськими моделями, що містять на кілька порядків більше параметрів.

- 4) У підрозділі 3.1.2 (с. 129-135), де наведено докладний опис наборів згенерованих даних від 15 різних моделей, бракує зведеної таблиці із зазначенням конкретних параметрів генерації, які використовувалися під час генерації, наприклад, temperature, top\_p та presence\_penalty для кожної з використаних архітектур. Наявність таких відомостей для Dense, MoE та Reasoning архітектур значно полегшила б можливість незалежного відтворення науковою спільнотою результатів експериментів.
- 5) Проводячи аналіз візуалізації результатів детекції у підрозділі 2.4.3 (Рис. 2.7 на с. 117), варто відзначити, що запропонований варіант карти доказів з деталізацією на рівні окремих tokenів є деталізованим та достатньо інформативним. Проте, у сценарії роботи з такими об'ємними документами, як, наприклад, дисертації чи журнали, експерту було б зручніше мати макрорівень перегляду матеріалу у вигляді теплової карти абзаців чи сторінок для прискорення процесу модерації.
- 6) Підрозділ 4.4.2 (с. 208-215) присвячено розбору прикладних кейсів використання ХАІ-інтерпретацій. Для повної та всебічної демонстрації можливостей запропонованих профілів атрибуції автор міг би проаналізувати змішаний текст, у якому одна половина написана людиною, а друга - згенерована ШІ. Це дозволило б наочно проілюструвати особливості поведінки системи та реакцію карти доказів безпосередньо на межі зміни авторства.

Наведені зауваження не знижують загального позитивного враження від дисертаційної роботи, а висвітлюють шляхи можливого подальшого розвитку проєкту.

## **8. Загальний висновок**

Дисертаційна робота Никоненка Андрія Олександровича на тему «Гібридний метод ідентифікації автоматично згенерованих природномовних текстів» є логічно викладеним, добре пропрацьованим, самостійним та актуальним науковим дослідженням. У ній отримано, висвітлено та обґрунтовано нові результати, а робота в цілому вирішує актуальну науково-прикладну задачу створення надійних систем детекції штучного інтелекту, підвищення їхньої точності, стійкості та зменшення рівня хибнопозитивних спрацювань.

За своєю актуальністю, рівнем наукової новизни, фундаментальністю, глибиною експериментальної валідації та вагомістю практичного застосування дисертація повністю відповідає вимогам Постанови Кабінету Міністрів України №44 від 12 січня 2022 року «Про затвердження Порядку присудження ступеня доктора філософії та скасування рішення разової спеціалізованої вченої ради закладу вищої освіти, наукової установи про присудження ступеня доктора філософії».

Автор дисертації, Никоненко Андрій Олександрович, заслуговує на присудження наукового ступеня доктора філософії за спеціальністю 122 «Комп'ютерні науки» галузі знань 12 «Інформаційні технології».

### Рецензент

д.т.н., професор,  
професор кафедри статистики та  
прикладної математики  
Черкаського державного  
технологічного університету



Євген ФЕДОРОВ

### Підпис

д.т.н., професора Є. Є. Федорова  
засвідчую  
Учений секретар ЧДТУ  
к.т.н., доцент



Ірина МИРОНЕЦЬ