

Облікова картка дисертації

I. Загальні відомості

Державний обліковий номер: 0826U005106

Особливі позначки: відкрита

Дата реєстрації: 03-09-2026

Статус: Запланована

Реквізити наказу МОН / наказу закладу:



II. Відомості про здобувача

Власне Прізвище Ім'я По-батькові:

1. Андрощук Максим Віталійович

2. Maksym Androshchuk

Кваліфікація:

Ідентифікатор ORCID ID: Не застосовується

Вид дисертації: доктор філософії

Аспірантура/Докторантура: так

Шифр наукової спеціальності: 122

Назва наукової спеціальності: Комп'ютерні науки

Галузь / галузі знань: інформаційні технології

Освітньо-наукова програма зі спеціальності: Комп'ютерні науки

Дата захисту: 05-09-2026

Спеціальність за освітою: Комп'ютерні науки

Місце роботи здобувача: Національний університет "Киево-Могилянська академія"

Код за ЄДРПОУ: 16459396

Місцезнаходження: вул. Г. Сковороди, Київ, 04070, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR: 03wfca816

III. Відомості про організацію, де відбувся захист

Шифр спеціалізованої вченої ради (разової спеціалізованої вченої ради): PhD 16411

Повне найменування юридичної особи: Національний університет "Києво-Могилянська академія"

Код за ЄДРПОУ: 16459396

Місцезнаходження: вул. Г. Сковороди, Київ, 04070, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR: 03wfca816

IV. Відомості про підприємство, установу, організацію, в якій було виконано дисертацію

Повне найменування юридичної особи: Національний університет "Києво-Могилянська академія"

Код за ЄДРПОУ: 16459396

Місцезнаходження: вул. Г. Сковороди, Київ, 04070, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR: 03wfca816

V. Відомості про дисертацію

Мова дисертації: Українська

Коди тематичних рубрик: 20.53.19, 28.23.39, 50.01.30

Тема дисертації:

1. Методи перевірки та безпечної актуалізації зовнішньої пам'яті RAG за умов модифікації знань та спотворення зворотного зв'язку
2. Methods for verification and reliable updating of external memory in RAG systems under knowledge modification and feedback distortion

Реферат/анотація:

1. Андрощук М. В. Методи перевірки та безпечної актуалізації зовнішньої пам'яті RAG за умов модифікації знань та спотворення зворотного зв'язку. – Кваліфікаційна наукова праця на правах рукопису. Дисертація на здобуття наукового ступеня доктора філософії за спеціальністю 122 «Комп'ютерні науки» (12 — Інформаційні технології). — Національний університет «Києво-Могилянська академія», Київ, 2026. Системи RAG дають змогу великим мовним моделям використовувати зовнішні джерела знань без перенавчання їхніх параметрів. У динамічних середовищах — корпоративних, медичних і юридичних — зовнішні джерела знань постійно змінюються, що породжує проблему безпечного оновлення: як приймати легітимні зміни й водночас відхиляти спотворені або помилкові дані. Більшість наявних RAG-систем не забезпечує формально обґрунтованих меж ризику проникнення отруєних даних, не має механізмів розрізнення природного дрейфу знань і координованого спотворення, а також потребує значних обчислювальних витрат для верифікації

суперечностей, які за повного попарного порівняння зростають квадратично. Наявні підходи до захисту RAG-систем, зокрема PoisonedRAG, RobustRAG і TrustRAG, розглядають окремі аспекти захисту — фільтрацію отриманих документів, агрегування відповідей із кількох підмножин та оцінювання довіри до джерел, однак не формують єдиної архітектури, яка поєднує верифікацію суперечностей, безпечне переведення нових знань із карантинного стану до основної пам'яті та розрізнення легітимного дрейфу знань і координованого отруєння. Окремою проблемою є обчислювальна вартість верифікації: повне попарне порівняння отриманих документів засобами NLI має квадратичну складність відносно кількості документів у контексті, що обмежує застосування такої верифікації в онлайн-сценаріях. Водночас параметричні підходи до редагування знань моделей, зокрема ROME і MEMIT, передбачають доступ до ваг або внутрішніх представлень базової моделі — це ускладнює їх застосування для закритих моделей, доступних лише через API і поширених у промислових RAG-розгортаннях. Об'єктом дослідження є процеси оброблення, верифікації та оновлення даних у зовнішній пам'яті RAG-систем за умов динамічної зміни знань і спотворення джерел інформації. Предметом дослідження є методи й алгоритми верифікації узгодженості, виявлення суперечностей та безпечної актуалізації зовнішньої пам'яті RAG-систем без модифікації параметрів генеративної моделі. Метою дослідження є розроблення методів верифікованої та захищеної від отруєння актуалізації зовнішньої пам'яті RAG-систем за умов обмеженої обчислювальної вартості верифікації. Для досягнення мети вирішено такі завдання: розроблено якірну схему виявлення суперечностей зі зниженою вартістю NLI-верифікації; запропоновано двофазну схему карантину та затвердження нових знань із модельною верхньою межею ризику проникнення отруєних даних; побудовано шар прийняття рішень для розрізнення легітимного дрейфу знань і координованого отруєння; удосконалено механізми запитно-орієнтованого відбору підграфа, злиття ранжувань і керування повторною генерацією. Методологічну основу становлять методи математичного моделювання, теорії ймовірностей, статистичного аналізу, оброблення природної мови (зокрема NLI), байєсівські методи виявлення зміни режиму та порівняльна експериментальна перевірка. Наукова новизна полягає у розробленні архітектури непараметричної адаптації пам'яті RAG, яка поєднує верифікацію, карантин і виявлення режиму змін у єдиному замкненому циклі. Завдяки спільній оцінці довіри до джерел, система накопичує дані та адаптується без ручного налаштування. Впроваджена якірна схема знизилася вартість NLI-верифікації з $O(K^2)$ до $O(K \cdot a)$. Експериментальна перевірка на 10 наборах ($N=1250$ запитів) та в умовах змагальних атак підтвердила зниження витрат на верифікацію на 57 %, блокування всіх отруєних документів та збереження 93 % легітимних оновлень. За цільової атаки відновлення якості відповідей склало 99,4 %. Конвеєр зберігає статистичний паритет із BM25-ретривером. Обмеження дослідження включають залежність від точності NLI-верифікатора, вразливість до масових Sybil-атак та необхідність валідації симульованого зворотного зв'язку на реальних користувачах. Практичне значення полягає у можливості побудови надійних систем (медичних, юридичних тощо) із закритою генеративною моделлю без її перенавчання. Ключові слова: машинне навчання, нейронні мережі, великі мовні моделі, інформаційний пошук, питально-відповідальні системи, пошукова генерація, RAG, графи знань, добування знань, верифікація знань, галюцинації, зовнішня пам'ять, актуалізація пам'яті, зворотний зв'язок, безпека даних.

2. Androshchuk M. V. Methods for verification and reliable updating of external memory in RAG systems under knowledge modification and feedback distortion. – Qualification scientific work submitted as a manuscript. PhD thesis to obtain the degree of Doctor of Philosophy in the Programme Subject Area 122 «Computer Science» (12 – Information technology). — National University of Kyiv-Mohyla Academy, Kyiv, 2026. Retrieval-Augmented Generation (RAG) systems in question-answering tasks allow large language models to use external knowledge sources without retraining their parameters. In dynamic environments—corporate, medical, and legal—external knowledge sources are constantly changing, which raises the problem of safe updating: how to accept legitimate changes while rejecting distorted or erroneous data. Most existing RAG systems do not provide formally justified bounds on the risk of poisoned data infiltration, lack mechanisms for distinguishing between natural knowledge drift and coordinated distortion, and require significant computational costs for contradiction verification, which grow quadratically under full pairwise comparison. Existing approaches to defending RAG systems, in particular

PoisonedRAG, RobustRAG, and TrustRAG, address individual aspects of defense—filtering retrieved documents, aggregating answers from multiple subsets, and evaluating source trust—but do not form a unified architecture that combines contradiction verification, the safe transfer of new knowledge from a quarantine state to main memory, and the distinction between legitimate knowledge drift and coordinated poisoning. A separate problem is the computational cost of verification: full pairwise comparison of retrieved documents using NLI (Natural Language Inference) has quadratic complexity relative to the number of documents in the context, which limits the application of such verification in online scenarios. At the same time, parametric approaches to model knowledge editing, such as ROME and MEMIT, require access to the weights or internal representations of the base model; this complicates their application to closed models accessible only via API, which are common in industrial RAG deployments. The object of the research is the processes of processing, verifying, and updating data in the external memory of RAG systems under dynamic knowledge changes and information source distortion. The subject of the research is the methods and algorithms for consistency verification, contradiction detection, and safe updating of the external memory of RAG systems without modifying the generative model's parameters. The purpose of the research is to develop methods for the verified and poison-protected updating of RAG external memory under conditions of limited computational verification costs. To achieve this purpose, the following tasks were accomplished: an anchor scheme for contradiction detection with reduced NLI verification costs was developed; a two-phase scheme for quarantining new knowledge with a modeled upper bound for poisoning risk was proposed; a decision-making layer for distinguishing legitimate knowledge drift from coordinated poisoning was built; and query-oriented subgraph selection and rank fusion mechanisms were improved. The methodological basis comprises mathematical modeling, probability theory, statistical analysis, natural language processing (specifically NLI), Bayesian methods for regime change detection, and comparative experimental validation. The scientific novelty lies in developing a non-parametric adaptation architecture for RAG external memory that combines verification, quarantine, and change regime detection in a closed loop. The shared evaluation of source trust allows the system to adapt without manual tuning. The introduced anchor scheme reduced the NLI verification cost from $O(K^2)$ to $O(K \cdot a)$. Experimental validation across 10 datasets ($N=1250$) and under adversarial attacks confirmed a 57% reduction in verification costs, the blocking of all poisoned documents, and the retention of 93% of legitimate updates. Under targeted attacks, response quality recovery reached 99.4%. The pipeline maintains statistical parity with a BM25 retriever. Limitations of the study include dependence on external NLI verifier accuracy, vulnerability to massive Sybil attacks, and the need to validate simulated feedback with real users. The practical significance lies in the ability to build reliable information systems compatible with closed APIs, without retraining language models. Keywords: machine learning, neural networks, large language models, information retrieval, question-answering systems, retrieval-augmented generation, RAG, knowledge graphs, knowledge extraction, knowledge verification, hallucinations, external memory, memory updating, feedback, data security.

Державний реєстраційний номер ДіР:

Пріоритетний напрям розвитку науки і техніки: Фундаментальні наукові дослідження з найбільш важливих проблем розвитку науково-технічного, соціально-економічного, суспільно-політичного, людського потенціалу для забезпечення конкурентоспроможності України у світі та сталого розвитку суспільства і держави

Стратегічний пріоритетний напрям інноваційної діяльності: Не застосовується

Підсумки дослідження: Нове вирішення актуального наукового завдання

Публікації:

- Андрощук, М., Кирієнко, О. (2020). Порівняння сервісів для створення питально-відповідальних систем. Наукові записки НаУКМА. Комп'ютерні науки, 3, 132–137.

- Андрощук, М. (2025). Сучасні підходи до використання баз знань для вирішення проблем великих мовних моделей. Наукові записки НаУКМА. Комп'ютерні науки, 7, 98–101.
- Андрощук, М. (2025). Вплив методів добування знань на ефективність RAG-систем на основі графів. Наукові записки НаУКМА. Комп'ютерні науки, 8, 84–87.
- Шевченко, М., Андрощук, М. (2025). Мультимодальний RAG з використанням текстових та візуальних даних. Проблеми програмування, (3), 66–78.
- Андрощук, М. (2024, 24 травня). Використання баз знань для підвищення точності та повноти результатів роботи великих мовних моделей. у Когнітивні дослідження: результати, виклики та перспективи: матеріали міжнародної науково-практичної конференції (с. 81–82). Київ, Україна.
- Androshchuk, M. (2024, 9–10 вересня). Efficiency of question-answering pipeline in systems with combined LLM and knowledge base usage. у Proceedings of the XXXIX International Conference on Problems of Decision Making Under Uncertainties (PDMU-2024) (с. 17). Brno, Czech Republic.
- Андрощук, М. (2025, 23–24 листопада). Проблеми фрагментації контексту в RAG системах / Problems of context fragmentation in RAG systems. у Теоретичні та прикладні аспекти побудови програмних систем (ТАAPSD 2025): праці 16-ї міжнародної науково-практичної конференції (с. 71–72). Київ, Україна.
- Андрощук, М. (2026, 7 травня). Зниження витрат на перевірку цілісності у системах генерації доповнених пошуком через якірний відбір. у Автоматизація та комп'ютерно-інтегровані технології – 2026 (AKIT – 2026): матеріали XII міжнародної науково-практичної конференції молодих учених, аспірантів і студентів (с. 110–111). Київ, Україна.

Наукова (науково-технічна) продукція:

Соціально-економічна спрямованість:

Охоронні документи на ОПІВ:

Впровадження результатів дисертації: Впроваджено

Зв'язок з науковими темами: 0125U002728

VI. Відомості про наукового керівника/керівників (консультанта)

Власне Прізвище Ім'я По-батькові:

1. Глибовець Андрій Миколайович

2. Andrii Hlybovets

Кваліфікація: д. т. н., професор, 01.05.03

Ідентифікатор ORCID ID: 0000-0003-4282-481X

Додаткова інформація:

Повне найменування юридичної особи: Національний університет "Кієво-Могилянська академія"

Код за ЄДРПОУ: 16459396

Місцезнаходження: вул. Г. Сковороди, Київ, 04070, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR: 03wfca816

VII. Відомості про офіційних опонентів та рецензентів

Офіційні опоненти

Власне Прізвище Ім'я По-батькові:

1. Жолткевич Григорій Миколайович
2. Grygoriy Zholtkevych

Кваліфікація: д. т. н., професор, 05.02.08

Ідентифікатор ORCID ID: 0000-0002-7515-2143

Додаткова інформація:

Повне найменування юридичної особи: Львівський національний університет імені Івана Франка

Код за ЄДРПОУ: 02070987

Місцезнаходження: вул. Університетська, Львів, 79000, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR:

Власне Прізвище Ім'я По-батькові:

1. Погорілий Сергій Дем'янович
2. Sergiy Pogorilyy

Кваліфікація: д. т. н., професор, 05.13.05

Ідентифікатор ORCID ID: 0000-0002-6497-5056

Додаткова інформація:

Повне найменування юридичної особи: Київський національний університет імені Тараса Шевченка

Код за ЄДРПОУ: 02070944

Місцезнаходження: вул. Володимирська, Київ, 01033, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR:

Рецензенти

Власне Прізвище Ім'я По-батькові:

1. Франчук Олег Васильович
2. Oleg Franchuk

Кваліфікація: к. т. н., доц., 05.13.06

Ідентифікатор ORCID ID: 0000-0002-1122-4689

Додаткова інформація:

Повне найменування юридичної особи: Національний університет "Кієво-Могилянська академія"

Код за ЄДРПОУ: 16459396

Місцезнаходження: вул. Г. Сковороди, Київ, 04070, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR: 03wfca816

Власне Прізвище Ім'я По-батькові:

1. Олецкий Олексій Віталійович

2. Oleksiy Oletsky

Кваліфікація: к. т. н., доц., 05.13.06

Ідентифікатор ORCID ID: 0000-0002-0553-5915

Додаткова інформація:

Повне найменування юридичної особи: Національний університет "Кієво-Могилянська академія"

Код за ЄДРПОУ: 16459396

Місцезнаходження: вул. Г. Сковороди, Київ, 04070, Україна

Форма власності: Державна

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR: 03wfca816

VIII. Заключні відомості

Власне Прізвище Ім'я По-батькові
голови ради

Малашонок Геннадій Іванович

Власне Прізвище Ім'я По-батькові
головуючого на засіданні

Малашонок Геннадій Іванович

Відповідальний за підготовку
облікових документів

Басенко Олена Едуардівна

Реєстратор

Юрченко Тетяна Анатоліївна

Керівник відділу УкрІНТЕІ, що є
відповідальним за реєстрацію наукової
діяльності



Юрченко Тетяна Анатоліївна