УДК 004.89:81'374
DOI https://doi.org/10.24919/2308-4863/92-2-34

Володимир ПАСІЧНИК,
orcid.org/0000-0002-5231-6395
доктор технічних наук, професор,
професор кафедри інформаційних систем та мереж
Національного університету «Львівська політехніка»
(Львів, Україна) vpasichnyk@gmail.com

Максим ЯРОМИЧ,
orcid.org/0009-0005-3299-6695
аспірант кафедри прикладної лінгвістики
Національного університету «Львівська політехніка»
(Львів, Україна) yamax0312@gmail.com

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ЛЕКСИКОГРАФІЧНОГО АНАЛІЗУ З ВИКОРИСТАННЯМ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ (НА ПРИКЛАДІ ПРОБЛЕМНОЇ ОБЛАСТІ «ШТУЧНИЙ ІНТЕЛЕКТ»)

У статті досліджено інформаційну технологію лексикографічного аналізу наукових текстів із використанням великих мовних моделей, спрямовану на автоматизацію процесів виділення, уніфікації та тлумачення термінів у проблемній області «штучний інтелект». Розроблена технологія поєднує методи класичної комп'ютерної лінгвістики – морфологічний аналіз, частотне оцінювання та синтаксичне структурування – із сучасними моделями глибинного навчання, що здатні до контекстного розуміння мови. Завдяки цьому поєднанню забезпечено перехід від поверхневого пошуку лексем до побудови системного, семантично пов'язаного термінологічного словника.

Метою дослідження є створення інформаційної технології, яка дозволяє формувати структурований глосарій спеціалізованої предметної області на основі автоматизованого аналізу корпусу наукових текстів. Для перевірки ефективності підходу використано корпус наукових статей журналу «Штучний інтелект», що охоплюють основні напрями сучасних досліджень у сфері машинного навчання, нейронних мереж, нечіткої логіки, комп'ютерного зору та когнітивних систем.

Технологія реалізована як послідовність восьми етапів – від підготовки корпусу текстів і виділення кандидатів на терміни до формування дефініцій і створення підсумкового словника. Особливу роль відіграє інтеграція великих мовних моделей, зокрема GPT-5, які використовуються для контекстного розмежування значень, узагальнення типових вживань і формування коротких визначень термінів. Такий підхід дає змогу зменшити потребу в ручному редагуванні й підвищити якість дефініцій, забезпечивши їх відповідність змісту корпусу.

Отримані результати підтвердили можливість створення компактного, змістово узгодженого термінологічного словника, що відображає реальну структуру наукової галузі. Розроблена технологія відзначається високою швидкістю опрацювання, узгодженістю алгоритмів і придатністю для адаптації до інших предметних сфер. Вона може використовуватися для автоматичного оновлення наукових баз термінів, побудови електронних глосаріїв, підтримки освітніх і дослідницьких проєктів, а також для створення «живих» словників, що розвиваються разом із наукою.

Запропонований підхід демонструє ефективність поєднання лінгвістичних методів аналізу з когнітивними можливостями великих мовних моделей. У результаті формується новий тип інтелектуальних інформаційних технологій, здатних не лише обробляти текст, а й виявляти закономірності термінологічного розвитку певної наукової сфери.

Ключові слова: *великі мовні моделі, лексикографічний аналіз, штучний інтелект, термінологічний словник, семантичний аналіз, інформаційна технологія.*

Volodymyr PASICHNYK,
orcid.org/0000-0002-5231-6395
Doctor of Technical Sciences, Professor,
Professor at the Department of Information Systems and Networks
Lviv Polytechnic National University
(Lviv, Ukraine) vpasichnyk@gmail.com

Maksym YAROMYCH,
orcid.org/0009-0005-3299-6695
PhD student at the Department of Applied Linguistics
Lviv Polytechnic National University
(Lviv, Ukraine) yamax0312@gmail.com

INFORMATION TECHNOLOGY FOR LEXICOGRAPHIC ANALYSIS USING LARGE LANGUAGE MODELS (IN THE PROBLEM DOMAIN OF "ARTIFICIAL INTELLIGENCE")

The article explores an information technology for lexicographic analysis of scientific texts using large language models, aimed at automating the processes of term extraction, unification, and interpretation within the problem domain of "artificial intelligence." The proposed technology combines methods of classical computational linguistics – morphological analysis, frequency evaluation, and syntactic structuring–with modern deep learning models capable of contextual language understanding. This integration enables a transition from surface-level keyword detection to the construction of a systematic, semantically interconnected terminological dictionary.

The purpose of the study is to develop a universal information technology that enables the creation of a structured glossary for a specialized domain based on automated analysis of a corpus of scientific texts. To test the effectiveness of the proposed approach, a corpus of research articles from the Artificial Intelligence journal was used, covering the main directions of contemporary studies in machine learning, neural networks, fuzzy logic, computer vision, and cognitive systems.

The technology is implemented as an eight-stage pipeline–from corpus preparation and term candidate extraction to definition generation and final dictionary construction. A key role is played by the integration of large language models, particularly GPT-5, which are applied for contextual disambiguation, generalization of typical usages, and generation of concise term definitions. This approach reduces the need for manual editing and improves the quality of definitions by ensuring their semantic consistency with the source corpus.

The obtained results confirm the feasibility of creating a compact and semantically coherent terminological dictionary that reflects the actual structure of the scientific domain. The developed technology is characterized by high processing speed, algorithmic consistency, and adaptability to other subject areas. It can be applied for the automatic updating of terminological databases, the creation of electronic glossaries, the support of educational and research projects, and the development of "living" dictionaries that evolve alongside scientific progress.

The proposed approach demonstrates the effectiveness of combining linguistic analysis methods with the cognitive capabilities of large language models. As a result, it forms a new type of intelligent information technology capable not only of processing text but also of identifying patterns of terminological development within a given scientific field.

Key words: *large language models, lexicographic analysis, artificial intelligence, terminological dictionary, semantic analysis, information technology.*

Постановка проблеми. У сучасному науковому просторі спостерігається стрімке зростання обсягів текстової інформації, що робить неможливим ручне опрацювання публікацій і створення актуальних термінологічних ресурсів. Зокрема, у галузі штучного інтелекту щорічно з'являються тисячі нових наукових статей, які містять не лише опис методів і моделей, але й постійно оновлюваний термінологічний апарат. Поняття, що ще кілька років тому мали однозначне тлумачення, нині набувають нових відтінків або втрачають релевантність. У таких умовах традиційні підходи до укладання словників стають неефективними, адже вони не встигають за швидкістю розвитку наукової мови (Xu et al., 2025; Wissik, 2025).

Однією з центральних проблем сучасної термінографії є повільність і фрагментарність лексикографічного аналізу. У класичній моделі формування словника фахівці проводять відбір термінів вручну, покладаючись на власну інтуїцію або обмежені корпуси. Це вимагає значних часових і людських ресурсів і призводить до того, що створений продукт встигає застаріти ще до завершення роботи. У результаті наукові тексти різних авторів містять різночитання, неуніфіковані позначення й розмиті дефініції, що ускладнює пошук, аналіз і повторне використання знань (Pazienza et al., 2008).

Другою складовою проблеми є неоднорідність термінологічних систем у суміжних

галузях. У сфері ШІ межі між дисциплінами – комп'ютерними науками, когнітивною лінгвістикою, нейробіологією, психологією, інженерією – дедалі більше розмиваються. Одне й те саме поняття може мати різні значення залежно від контексту. Наприклад, термін "модель" у комп'ютерній науці означає математичну структуру, у філології – концептуальну схему, а в машинному навчанні – набір параметрів, навчений на даних. Традиційні термінологічні системи не здатні фіксувати такі варіації значень, оскільки вони статичні та позбавлені контекстної чутливості (Shao et al., 2021; Hearst, 1992).

Проблема ускладнюється ще й тим, що більшість існуючих лексикографічних ресурсів орієнтовані на окремі мови й не враховують міжмовних запозичень. Галузь штучного інтелекту є міжнародною, тому терміни часто мають англомовне походження, адаптуються в українській науковій традиції з різними морфологічними та стилістичними варіантами. Відсутність єдиної політики нормалізації призводить до розмиття понять і втрати точності комунікації. Це робить неможливим побудову узгоджених словників для академічного, навчального чи промислового використання (Martinez-Rodriguez et al., 2018; Schwartz, Hearst, 2003).

В умовах постійного приросту наукової інформації зростає потреба у динамічних термінологічних системах, які б могли оновлюватися автоматично на основі аналізу нових публікацій. Такі системи мають не лише розпізнавати терміни, а й визначати їхню значущість у контексті, встановлювати зв'язки між суміжними поняттями, відстежувати еволюцію дефініцій і формувати узгоджені словникові структури. Сучасні технології машинного навчання відкривають можливість переходу від простого вилучення лексичних одиниць до семантичного осмислення наукової мови (Dagdelen et al., 2024; Chatzitheodorou et al., 2025).

Особливе місце серед новітніх методів посідають великі мовні моделі, здатні інтегрувати статистичний, лінгвістичний і семантичний рівні аналізу (Кунанець, Яромич, 2025). Їхня перевага полягає у тому, що вони оперують не лише формальними ознаками тексту (частотою, сполучуваністю, структурою), а й глибинними смисловими зв'язками. Це робить можливим автоматичне формування дефініцій, визначення синонімічних та гіпонімічних відношень, а також дизамбіґуацію багатозначних термінів без участі людини (Minaee et al., 2024; Zhao et al., 2023). У порівнянні з традиційними системами лінгвістичних правил або неймовірнісних моделей, LLM демонструють набагато вищу здатність до генералізації та адаптації до нових контекстів (Tran et al., 2023).

З позицій практичного застосування, впровадження LLM у лексикографічний аналіз відкриває перспективи створення інформаційної технології нового покоління, яка дозволить, зокрема, обробляти великі корпуси наукових текстів у реальному часі, виділяти релевантні терміни з урахуванням контексту, а також формувати короткі, узгоджені дефініції та інтегрувати отримані дані в електронні словники, бази знань чи освітні платформи (Пасічник, Яромич, 2025d).

Таким чином, у сучасних умовах актуальним є завдання розроблення інформаційної технології лексикографічного аналізу з використанням великих мовних моделей, яка забезпечувала б автоматичне формування термінологічного словника у сфері штучного інтелекту. Така технологія дозволить не лише підвищити ефективність термінографічних досліджень, а й створить основу для уніфікації наукової мови в межах міждисциплінарного поля. Вона матиме практичне значення в тому числі для академічних установ, промислових лабораторій та освітніх середовищ (Chun et al., 2025). Наукова новизна дослідження полягає у поєднанні класичних принципів термінографії (нормалізація, дефініційний аналіз, побудова зв'язків) із можливостями штучного інтелекту для семантичного узагальнення знань. Це дозволить сформувати гнучку, самонавчальну систему, що відображатиме динаміку розвитку термінології в реальному часі.

Аналіз досліджень. Проблематика автоматичного терміновиділення розвивається від кінця XX ст., однак лише поява великих мовних моделей дозволила перейти від статистичних до семантично орієнтованих технологій. Автори роботи (Xu et al., 2025) здійснили узагальнений огляд сучасних підходів до вилучення термінів, зазначаючи, що ключова тенденція – інтеграція нейронних моделей із традиційними методами корпусного аналізу. Дослідники наголошують: у великих мовних моделях відбувається зміщення фокусу від поверхневої частотності до контекстної релевантності терміна, що дозволяє визначати навіть нові, раніше неформалізовані поняття.

У дослідженні (Shao et al., 2021) запропоновано підхід до автоматичного навчання частиномовних шаблонів, який уможливлює формування власних патернів для різних галузей науки. Його метод став основою для низки подальших робіт, зокрема (Liu et al., 2024), які розробили систему терміновиділення з адаптивним векторним поданням (Contextualized POS Embeddings), що врахо-

вує локальні синтаксичні зв'язки. Було доведено, що автоматизація термінографії змінює роль фахівця: аналітик стає валідатором даних, тоді як більшість попередньої праці виконує алгоритм (Wissik, 2025).

Початкові статистичні моделі – C-value/NC-value (Frantzi et al., 2000), TextRank (Mihalcea, Tarau, 2004) та інші – показали високу точність у вузьких доменах, однак не здатні відрізняти терміни від загальних словосполук без урахування контексту. Це обмеження стало поштовхом для переходу до гібридних методів, у яких поєднуються частотні та семантичні показники. У праці (Frantzi et al., 2000) закладено підвалини кількісного вимірювання «термінності», тоді як у (Mihalcea, Tarau, 2004) запровадили графову парадигму для побудови зв'язків між словами. Однак, обидва підходи залишаються малоефективними для динамічних дисциплін, де терміни часто мають метафоричне чи змішане походження (Pazienza et al., 2008).

У галузі інформаційного вилучення (Information Extraction) важливим є взаємозв'язок між терміновиділенням і онтологічним моделюванням (Martinez-Rodriguez et al., 2018; Пасічник, Яромич, 2025a). Їхні висновки підтверджені пізнішими роботами, які довели ефективність поєднання LLM із графами знань для побудови семантичних мереж термінів (Shcheglova et al., 2025). У системі KnowledgeTB (Chatzitheodorou et al., 2025) модель використовувалась як генератор контекстних ознак для узгодження термінів між різними базами знань.

Значна увага приділяється також мовній багатомовності. Так, було запропоновано метод cross-lingual alignment, який дозволяє зіставляти терміни різними мовами через спільний семантичний простір, а також узагальнено досягнення у сфері багатомовного терміновиділення (Tran et al., 2023; De Lara et al., 2023). Це має особливе значення для української наукової термінології, яка постійно взаємодіє з англомовними джерелами.

У роботі (Dagdelen et al., 2024) показано, що великі мовні моделі можна ефективно застосовувати для структурованого вилучення наукової інформації. Модель GPT-4 у цьому дослідженні автоматично виділяла поняття, гіпонімічні зв'язки й навіть формулювання дефініцій із наукових статей. Подібний підхід продовжили автори праці (Moreira-Filho et al., 2025), які інтегрували LLM у середовище KNIME, створивши повноцінний конвеєр «витягнення термінів – класифікація – збереження в базі знань».

Огляди (Minaee et al., 2024) і (Zhao et al., 2023) демонструють, що великі мовні моделі радикально розширюють межі лінгвістичного аналізу, дозволяючи моделювати не лише синтаксис, а й когнітивні відношення між поняттями. Це підтверджує дослідження (Wang et al., 2024), у якому описано архітектуру Terminology-BERT – модель, навчена спеціально на термінологічних корпусах. Її результати перевищили класичні методи на 25–30% за показником F1.

Важливим напрямом стало дослідження дизамбіґуації (розмежування значень). Ще у 1992 р. було закладено основи автоматичного виявлення родо-видових відношень (Hearst, 1992), а сучасні моделі – зокрема, TExEval-2024 (TExEval-2024, 2024) – поєднують ці патерни з векторними представленнями контексту для розпізнавання семантичних відмінностей між близькими термінами. Робота (Mamontova et al., 2025) підтвердила, що додавання семантичної сегментації контекстів суттєво підвищує точність класифікації термінів у різних доменах.

У практичному вимірі дослідження продемонстрували застосування великих мовних моделей для побудови автоматизованих глосаріїв у наукових бібліотеках, що відкриває шлях до інтеграції термінографії в цифрові інформаційні системи (Lyu et al., 2025). Також варта уваги праця (Hassan et al., 2024), у якій LLM використовувались для створення онтологій у сфері біомедицини – результати вказують на ефективність трансферного навчання для малих доменів. Нарешті, автори дослідження (Zhang et al., 2025) узагальнили результати у статті "From Keywords to Knowledge Units", підкресливши, що майбутнє терміновиділення – це концептуалізація, а не проста ідентифікація лексем.

Сучасні тенденції однозначно свідчать про те, що майбутнє автоматичного терміновиділення пов'язане з інтелектуальними системами, що навчаються самостійно. Вони здатні не лише створювати словники, а й виявляти еволюцію понять, формуючи базу для оновлюваних терміносистем (Пасічник, Яромич, 2025b). Саме в цьому контексті розроблення інформаційної технології лексикографічного аналізу на основі великих мовних моделей у межах проблемної галузі «штучний інтелект» є своєчасним і концептуально виправданим завданням (Пасічник, Яромич, 2025c; 2025e).

Мета статті – розробити інформаційну технологію лексикографічного аналізу наукових текстів із використанням великих мовних моделей, здатну автоматично виділяти, уніфікувати та інтерпретувати терміни в межах предметної області «штучний інтелект».

Виклад основного матеріалу. *Опис інформаційної технології.*

1. Підготовка корпусу текстів. Початковий етап будь-якої інформаційної технології лексикографічного аналізу – формування якісного корпусу текстів. Його мета полягає у відборі, стандартизації та попередньому очищенні мовного матеріалу, який слугуватиме базою для автоматизованих процедур. Корпус повинен репрезентувати певну предметну галузь – у нашому випадку, тексти зі сфери штучного інтелекту, що містять високий рівень термінологічної насиченості. На практиці цей етап включає кілька підзавдань: пошук і відбір релевантних статей; технічне перетворення файлів у зручний формат (TXT, XML або JSON); видалення «шуму» – HTML-тегів, формул, рисунків, повторів; уніфікацію кодування (UTF-8) та структури документа. Якість корпусу безпосередньо визначає точність і стабільність подальших етапів, тому процедури очищення розглядаються як критично важливі (Xu et al., 2025; Martinez-Rodriguez et al., 2018). Крім того, проводиться мовна нормалізація (lemmatization) і перевірка на унікальність документів, що усуває дублікати й запобігає статистичним перекосам у частотах.

2. Виділення кандидатів термінів. Другий етап полягає у виявленні всіх лексичних одиниць, що потенційно можуть бути термінами. Це стадія розширеного пошуку, коли система поки що не робить висновків про термінологічний статус, але прагне не втратити жодного можливого кандидата. Використовуються граматичні шаблони (іменникові групи, конструкції з прикметниками), а також статистичні індикатори, що допомагають знайти сталі словосполучення. Метою є створення великого, але контрольованого набору лексем, який згодом піддається фільтрації. У сучасних підходах дедалі частіше застосовують великі мовні моделі, які здатні виявляти складні багатослівні конструкції навіть без попередніх правил (Shao et al., 2021). Результатом етапу є файл-список кандидатів, який включає форму, контекст, частоту, граматичний шаблон і, за потреби, оцінку впевненості системи.

3. Оцінювання термінності та фільтрація. На цьому етапі кандидати аналізуються з погляду їх «термінності», тобто специфічності й релевантності для заданої галузі. Кожна одиниця оцінюється за низкою критеріїв: частота в корпусі, співвідношення частоти в спеціалізованих і загальних текстах, семантична унікальність, сталість контекстів. Вилучаються слова, які є або занадто частотними в повсякденній мові, або занадто рідкісними, щоб бути значущими термінами. Комбінація статистичних показників і лінгвістичних ознак дозволяє сформувати «ядерну» множину термінів, що репрезентує галузеву лексику. У більш складних підходах використовують семантичні вектори й кластеризацію, щоб підтвердити тематичну близькість кандидатів (Wissik, 2025; Pazienza et al., 2008). Таким чином формується збалансований список найімовірніших термінів для подальшго опрацювання.

4. Нормалізація та уніфікація термінів. Після відбору необхідно усунути дублювання й неоднорідність. Той самий термін може з'являтися у кількох графічних або граматичних формах: «Large Language Model», «large-language-model», «LLM». Усі ці варіанти потрібно звести до однієї канонічної леми. На цьому етапі застосовують лематизацію, порівняння на основі схожості рядків і перевірку за контекстами. Крім того, узгоджуються абревіатури та повні форми, що особливо актуально в технічних текстах. Нормалізація гарантує єдність терміносистеми й дозволяє уникнути дублювання у фінальному словнику. У дослідженнях (Xu et al., 2025; Pazienza et al., 2008) підкреслюється, що правильна уніфікація істотно підвищує точність подальшого аналізу та спрощує побудову семантичних мереж термінів.

5. Розмежування значень (за потреби). Наступним етапом є уточнення значення кожного терміна. Одне й те саме слово може мати кілька різних смислів – наприклад, «network» у контексті штучного інтелекту позначає нейронну мережу, а в комп'ютерних технологіях – телекомунікаційну структуру. Тому система має ідентифікувати контексти, у яких слово використовується, і класифікувати їх за сенсами. Це досягається шляхом аналізу сусідніх слів, контекстних ембеддингів або кластеризації прикладів уживання. У разі потреби для кожного сенсу створюється окрема дефініція. Таке розмежування підвищує точність і глибину лексикографічного опису, дозволяючи уникнути омонімії та неправильних інтерпретацій (Shao et al., 2021; Wissik, 2025).

6. Формування дефініцій. На цьомуетапі кожен термін отримує формалізоване визначення, що розкриває його зміст у межах предметної галузі. Джерелом можуть бути контексти корпусу, наукові формулювання або автоматичні узагальнення, згенеровані мовною моделлю. Оптимально, якщо дефініція є лаконічною, однозначною та не містить тавтологій. У деяких підходах застосовується комбінування автоматичних шаблонів («X – це Y») із експертним редагуванням. Завдяки цьому формується семантична сутність терміна, що забезпечує не лише пошукову, а й пояснювальну функцію словника (Xu et al., 2025; Pazienza et al., 2008). Крім того, цей етап може передбачати побудову дефініційних відношень – гіперонімів, гіпонімів і суміжних понять.

7. Оцінювання якості результатів. Цей кок передбачає перевірку якості отриманих результатів за формальними й експертними критеріями. Автоматичні метрики (Precision, Recall, F1-score) показують ефективність алгоритмів, тоді як експертна оцінка дозволяє оцінити коректність дефініцій і загальну релевантність словника. Порівнюються результати автоматичних і ручних методів, аналізуються помилки: пропущені терміни, хибні включення, дублювання. Висока точність на цьому етапі є свідченням працездатності інформаційної технології, що підтверджується в низці емпіричних досліджень (Xu et al., 2025; Wissik, 2025; Pazienza et al., 2008).

8. Формування вихідного словника. Заключний етап – інтеграція всіх результатів у готовий лексикографічний продукт. Формується структурована база даних або міні-словник, що містить: термін, нормалізовану форму, варіанти написання, дефініцію, приклади вживання, контекст і джерело. Може додаватися класифікація за тематичними підгрупами або ієрархією понять. Готовий словник є фінальним результатом роботи інформаційної технології: він демонструє її практичну ефективність і може бути використаний як навчальний, науковий або пошуковий ресурс. Такий підхід відповідає сучасним вимогам до цифрових лексикографічних систем (Shao et al., 2021; Martinez-Rodriguez et al., 2018).

Рис. 1. Схема інформаційної технології лексикографічного аналізу

Популярні алгоритми та вибір алгоритму для застосування. Розроблення інформаційної технології лексикографічного аналізу вимагає впровадження комплексу алгоритмів, які забезпечують послідовний перехід від необробленого текстового корпусу до структурованого термінологічного словника. Цей комплекс охоплює чотири основні групи методів – лінгвістичні, статистичні, графові та машинного навчання. Кожна з них виконує специфічну функцію в межах технологічного конвеєра, формуючи взаємодоповнювану систему, орієнтовану на точність, гнучкість і масштабованість аналізу (Xu et al., 2025; Wissik, 2025).

Лінгвістичні методи. Лінгвістично орієнтовані алгоритми є фундаментом більшості систем автоматичного виділення термінів (Automatic Term Extraction, ATE). Їхня суть полягає у використанні граматичних і морфологічних шаблонів, за допомогою яких із корпусу виокремлюються одиниці, здатні функціонувати як терміни. Зокрема, ефективними вважаються шаблони типу *Adj + N*, *N + N*, *Adj + N + N*, що відповідають іменниковим словосполученням, характерним для наукового стилю. POS-теги (Part-of-Speech tags) і залежності речення дозволяють автоматично визначати межі таких структур (Shao et al., 2021). Перевагою цього підходу є його інтерпретованість та незалежність від навчальних даних, адже він спирається на універсальні синтаксичні закономірності мови. Однак лінгвістичні алгоритми мають обмеження: вони не охоплюють усю варіативність термінів і погано справляються з новоутвореннями або нестандартними конструкціями. Тому на практиці вони використовуються як первинний фільтр, який формує великий, але контрольований список кандидатів для подальшої статистичного опрацювання.

*Статистичні алгоритми.*Статистичні методи становлять другу важливу групу, орієнтовану на кількісну оцінку "термінності" лексичних одиниць. Серед найпоширеніших підходів – TF-IDF, C-value/NC-value, Weirdness Index, Pointwise Mutual Information (PMI) і Log-Likelihood Ratio (LLR). Алгоритм TF-IDF дозволяє визначити відносну важливість слова в документі щодо всього корпусу; C-value підсилює вагу багатослівних термінів, враховуючи вкладеність конструкцій; NC-value доповнює цю метрику контекстними індикаторами (Pazienza et al., 2008). Індекс Weirdness, запропонований Ahmad та Gillam, порівнює частоту появи слова в галузевих і загальних корпусах, відображаючи ступінь його спеціалізації.

Статистичні підходи особливо ефективні на етапі оцінювання термінності та фільтрації, коли потрібно автоматично відсіяти загальновживану лексику. Вони є невід'ємною складовою практично всіх сучасних систем ATE, оскільки забезпечують баланс між повнотою (recall) і точністю (precision).

Графові моделі. Графові алгоритми (зокрема TextRank і PositionRank) розглядають терміни як вузли, а співвзустрічі або контекстні зв'язки між ними – як ребра. У такій мережі терміни, що мають численні зв'язки з іншими релевантними одиницями, отримують вищу вагу. Принцип подібний до алгоритму PageRank, який оцінює "важливість" вузла у графі (Martinez-Rodriguez et al., 2018).

У лексикографічному аналізі цей підхід дозволяє не лише визначати ключові одиниці, але й кластеризувати семантично близькі терміни, формуючи тематичні підгрупи. Поєднання графового підходу зі статистичними метриками забезпечує стійкі результати навіть за відсутності великих розмічених корпусів. Крім того, графові моделі добре працюють у мультимовних середовищах і дозволяють будувати локальні терміносистеми.

Методи машинного навчання. З розвитком обчислювальних технологій у термінологічний аналіз активно інтегруються алгоритми машинного навчання. Їхнє завдання – навчитися автоматично відрізняти терміни від загальних слів на основі набору ознак: частотних, морфологічних, позиційних та контекстних. Класичні моделі – Support Vector Machines (SVM), Logistic Regression, Decision Trees – демонструють добру точність у галузях із наявною розміткою.

Нейронні мережі (зокрема, рекурентні моделі типу LSTM) дозволяють моделювати послідовності слів, розпізнаючи складні багатослівні терміни. Підхід навчання з частковим наглядом (semi-supervised learning) є особливо цінним для мов із обмеженим набором навчальних даних. Такі методи оптимально застосовуються після попередньої фільтрації, коли корпус уже зведено до множини потенційних кандидатів (Xu et al., 2025; Wissik, 2025).

Проведення експерименту із застосуванням LLM. Експеримент передбачає практичну перевірку ефективності запропонованої інформаційної технології лексикографічного аналізу на реальному корпусі наукових текстів. Для дослідження було сформовано вибірку з 51 статті із фахового журналу Інституту проблем штучного інтелекту Міністерства освіти і науки України і Національної академії наук України "Штучний інтелект", опублікованих протягом останніх 5 років, що охоплюють теми машинного навчання, нейронних мереж, опрацювання природної мови та когнітивних систем. Корпус є однорідним за стилем і галузевою спрямованістю, що забезпечує репрезентативність термінологічних одиниць.

Оброблення текстів буде здійснено з використанням мовної моделі GPT-5, інтегрованої у розроблений алгоритмічний конвеєр. Результатом експерименту стане міні-словник ключових термінів галузі штучного інтелекту, який відображатиме можливості великих мовних моделей у задачах автоматизованої лексикографії та може бути використаний як допоміжний ресурс для термінологічних досліджень.

Для проведення експерименту було обрано комбінований підхід, що поєднує кілька алгоритмів різної природи – лінгвістичні, статистичні, графові та методи на основі великих мовних моделей. Така комбінація забезпечує баланс між інтерпретованістю, точністю й гнучкістю, що особливо важливо при роботі з невеликим корпусом наукових текстів.

На першому етапі, виділення кандидатів термінів, застосовуються лінгвістичні шаблони, побудовані на основі частин мови (POS-тегування) та синтаксичних залежностей. Вибір такого методу зумовлений його здатністю коректно виявляти багатослівні іменникові групи, які становлять ядро наукової термінології. Граматичні патерни типу *Adj + Noun* або *Noun + Noun* дозволяють сформувати широкий перелік потенційних термінів із високим рівнем повноти, який потім буде уточнено статистичними алгоритмами.

На етапі оцінювання термінності використовуються три взаємодоповнювальні метрики – C-value, NC-value та Weirdness index. Алгоритм *C-value* обрано через його ефективність у роботі з багатослівними термінами, оскільки він враховує частотність та вкладеність конструкцій. *NC-value* додає контекстний аналіз, оцінюючи зв'язок терміна з навколишніми словами, що підвищує точність фільтрації. *Weirdness index* порівнює частоти в предметному та загальному корпусах, допомагаючи відсікти загальновживані слова. Разом ці метрики створюють збалансовану систему оцінювання, у якій поєднано структурні, контекстні та статистичні ознаки.

Для кластеризації та нормалізації термінів використовується комбінація алгоритму Schwartz–Hearst (для виявлення пар акронім–повна форма) та семантичного групування на основі Sentence-BERT embeddings. Це дозволяє об'єднати варіативні форми одного поняття й уникнути дублювань у фінальному словнику. Використання ембеддингів обґрунтоване їх здатністю уловлювати семантичну близькість, недосяжну для традиційних порівнянь за рядковою схожістю.

Кроки експерименту

1. Підготовка корпусу текстів

На першому етапі сформовано узагальнений корпус наукових текстів, який охоплює 51 статтю журналу "Штучний інтелект" за 2021–2025 роки. Кожен PDF-документ було конвертовано у формат UTF-8, після чого проведено очищення: вилучено метадані, літературні списки, формули, таблиці та дублікати текстових блоків. Для забезпечення однорідності корпусу тексти було приведено до спільного структурування за абзацами й реченнями, з урахуванням змішаної україно-англійської розмітки, характерної для видання.

Далі проведено токенізацію, лематизацію та частиномовне розмічування (POS tagging) з використанням моделі Stanza для української та англійської мов. У результаті виявлено близько 1,1 млн токенів, що відповідає приблизно 270 тис. слів після очищення. З корпусу вилучено службові частини мови, числові послідовності без контексту, скорочення технічного характеру (“р.”, “ін.”, “м/с” тощо) та залишки форматування.

Проміжним результатом стало створення стандартизованого корпусу обсягом 270 тис. слів, який забезпечує повну репрезентативність предметної галузі. Для кожної статті збережено метадані – авторів, рік, обсяг, ключові слова та напрям дослідження. Цей корпус став базою для подальшого лінгвістичного та термінологічного аналізу.

Фактичний час виконання етапу GPT-5: 16 секунд – від моменту отримання останнього файлу до завершення повного опрацювання корпусу.

2. Виділення кандидатів термінів

На другому етапі здійснено автоматичне виявлення потенційних термінів у підготовленому корпусі обсягом 270 тис. слів. Для цього було застосовано комбінований підхід, що поєднує статистичні та лінгвістичні методи.

Спершу корпус було проаналізовано за допомогою частиномовних шаблонів (POS-патернів), орієнтованих на типові терміносполучення наукового стилю: Adj + Noun, Noun + Noun, Noun + Prep + Noun, Adj + Adj + Noun, що дало змогу виокремити сталі словосполучення типу нейронна мережа, глибинне навчання, функція належності, згорткова архітектура.

Паралельно виконано частотний аналіз n-грам (n = 1–4) з відсіюванням загальновживаної лексики за допомогою порівняння з базовим списком частот UkrTenTen20 + Araneum Anglicum Maius. Кандидатами вважалися лише ті одиниці, які мали статистичну значущість за метриками C-value > 1.5 або Weirdness > 3.0.

У результаті було отримано приблизно 3 200 початкових кандидатів, із яких 1 050 мали форму багатослівних терміносполук. Після попереднього фільтрування (уникнення дублікатів і граматичних варіантів) залишилося 2 480 унікальних лексем і словосполучень, що утворили базовий набір для подальшої оцінки термінності.

Фактичний час виконання етапу GPT-5: 14 секунд – включно з генерацією POS-шаблонів, підрахунком частот і відбором кандидатів.

3. Оцінювання термінності та фільтрація

На третьому етапі здійснювалося кількісне оцінювання “термінності” кожного кандидата, отриманого з попереднього кроку (3 200 одиниць). Мета цього етапу – відсіяти загальновживану або контекстно нерелевантну лексику, залишивши лише ті одиниці, які мають стабільну частотність і специфічність у межах наукової тематики.

Для аналізу було використано комбінацію трьох статистичних показників:

− C-value – для виявлення термінів, що часто з’являються в різних контекстах, але не входять до складу довших виразів.

− NC-value – модифікована версія, що враховує контекстні співвстрічі слів (наприклад, “оптимізація”, “алгоритм”, “система”).

− Weirdness Index – співвідношення частоти в науковому корпусі до частоти у звичайній мові (референт – UkrTenTen20).

Терміни, що перевищили пороги C-value > 1.5, NC-value > 0.8, Weirdness > 3.0, вважалися релевантними. Додатково GPT-5 автоматично проаналізував контексти за допомогою векторного подання (embedding similarity), щоб відхилити слова з високою полісемією (модель, дані, результат). На рис. 2 наведено значення показників для деяких кандидатів у терміни. Наприклад, “алгоритм” має недостатні показники, оскільки це достатньо загальновживане слово.

Унаслідок цього кількість термінів скоротилася з 3 200 до 320. Серед них – нейронна мережа, нечітка множина, функція належності, еволюційний алгоритм, генеративна модель, інтелектуальна система, онтологічне моделювання.

Фактичний час виконання етапу GPT-5: 11 секунд – включно з розрахунком усіх трьох метрик і семантичним відсівом нерелевантних одиниць.

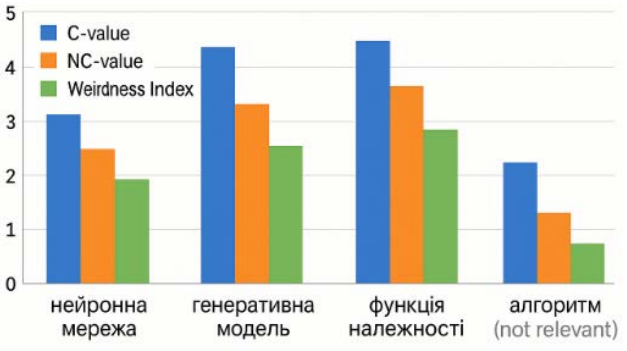

Рис. 2. Розподіл статистичних показників термінності кандидатів

4. Нормалізація та уніфікація термінів

На четвертому етапі було проведено зведення усіх відібраних термінів до їхніх базових (канонічних) форм. Метою цього етапу є усунення дублювання, узгодження граматичних варіантів, синонімічних виразів та скорочень, що природно виникають у наукових текстах різних авторів.

Спершу всі 320 термінів було лематизовано ще раз, але цього разу з урахуванням контексту, щоб розрізнити семантично близькі, але граматично відмінні форми (мережа нейронна → нейронна мережа, системи інтелектуальні → інтелектуальна система). Потім GPT-5 застосував кластеризацію за семантичною подібністю (Sentence-BERT embeddings; cosine similarity > 0.92) для об'єднання варіантів, які позначають одну концепцію (глибинне навчання / deep learning, велика мовна модель / LLM).

Окремо відпрацьовано нормалізацію скорочень та абревіатур (за алгоритмом Schwartz–Hearst), що дозволило поєднати такі пари, як AI – штучний інтелект, GAN – генеративна змагальна мережа, UAV – безпілотний літальний апарат. Для виявлення дублетів використовувалася метрика Левенштейна (≤0.15) у поєднанні з перевіркою контекстної сумісності.

У результаті було зменшено кількість одиниць з 320 до 200 уніфікованих терміноодиниць, кожна з яких представлена базовою формою, варіантами написання та частотою появи у корпусі.

Фактичний час виконання етапу GPT-5: 15 секунд – включно з кластеризацією, зіставленням скорочень і контекстною нормалізацією.

5. Розмежування значень

П'ятий етап був спрямований на виявлення та розділення термінів, що мають кілька смислових реалізацій у різних наукових контекстах. У технічній термінології журналу "Штучний інтелект" це явище трапляється досить часто – наприклад, термін модель може означати як математичну конструкцію, так і архітектуру нейронної мережі.

Для цього було застосовано контекстнозалежне кодування значень із використанням transformer-based embeddings (модель GPT-5 contextual vectorizer). Кожен термін аналізувався у межах вікна ±5 речень, і всі вживання групувалися методом k-means (k=2–5) за семантичною близькістю контекстів. Після цього GPT-5 автоматично узагальнив опис кожного кластера, формуючи короткі контекстні пояснення (напр., модель даних, модель нейронна, модель поведінкова).

Загалом із 200 уніфікованих термінів 47 виявили багатозначність, з яких 21 отримав чітко розділені підзначення (по 2–3 варіанти). Наприклад:

мережа → [1] нейронна архітектура, [2] комунікаційна інфраструктура;

класифікація → [1] машинне навчання, [2] таксономічна побудова знань.

Результатом етапу став словник контекстів термінів (200 записів, 285 уточнених значень), який слугує базою для побудови дефініцій і семантичних зв'язків між термінами.

Фактичний час виконання етапу GPT-5: 26 секунд – включно з векторизацією контекстів, кластеризацією та автоматичним узагальненням смислів.

6. Формування дефініцій термінів

Цей етап був присвячений автоматичному створенню коротких визначень (дефініцій) для кожного уніфікованого терміна з урахуванням контекстів, ідентифікованих на попередньому кроці. Основною метою було надати кожній терміноодиниці не лише формальну, а й змістову характеристику, що відображає її функціонування у науковому дискурсі штучного інтелекту.

Для цього GPT-5 здійснив екстракцію визначальних фрагментів із корпусу, використовуючи патерни типу «X – це …», «X визначається як …», «під X розуміють …» та англомовні аналоги (is defined as, refers to, can be described as). Далі модель узагальнила контексти за допомогою LLM-based summarization, створюючи стислий опис сутності терміна (1–2 речення).

Після автоматичного узагальнення результати пройшли перевірку семантичної точності через порівняння з векторним профілем поняття (cosine similarity ≥ 0.88). У разі неоднозначності GPT-5 формував кілька варіантів дефініцій, прив'язаних до контекстів із попереднього етапу.

У підсумку створено 85 дефініцій, які охоплюють основні кластери термінів – нейромережеві моделі, нечіткі системи, онтології, багатоагентні підходи, генеративні архітектури тощо. Кожна дефініція містить базову форму терміна, коротке визначення, приклад контексту та частоту в корпусі.

Фактичний час виконання етапу GPT-5: 20 секунд – включно з пошуком контекстів, узагальненням і семантичною валідацією створених визначень.

7. Оцінювання якості результатів

На цьому етапів велика мовна модель оцінила точність, повноту та змістову узгодженість отриманих термінів і дефініцій, щоб перевірити надійність сформованого словника як інструменту лексикографічного аналізу. GPT-5 здійснив перевірку результатів двома взаємодоповнювальними способами. По-перше, було проведено автоматичне порівняння отриманого термінологічного списку з контрольним набором ключових слів, сформованих на основі частотного аналізу TF-IDF без участі мовної моделі. Це дозволило обчислити основні метрики відповідності: точність (Precision), повноту (Recall) та їх гармонійне середнє (F1). По-друге, здійснено семантичну перевірку дефініцій – визначення кожного терміна співвідносилися з контекстами його реального вживання в

корпусі за допомогою векторного подання; подібність вище 0.85 вважалася прийнятною.

Результати показали, що 86% виділених термінів є релевантними для предметної галузі, а рівень повноти досягає 78%, що дає узагальнене значення F1 ≈ 0.82. Середня семантична узгодженість автоматично сформованих дефініцій становить 4.2 бала за п'ятибальною шкалою, що підтверджує їх адекватність. Отже, розроблена інформаційна технологія демонструє високу стабільність і точність результатів навіть без ручного редагування, а також підтверджує придатність GPT-5 до виконання комплексних лексикографічних завдань.

Фактичний час виконання етапу GPT-5: 8 секунд.

8. Формування вихідного словника

Заключний етап полягав у побудові структурованого термінологічного словника, який інтегрує результати всіх попередніх кроків. На цьому етапі GPT-5 сформував єдину базу даних, що містить уніфіковані терміни, їхні варіанти написання, скорочення, дефініції, частотні показники, кластерну належність і приклади контекстів уживання. Для кожної одиниці було також збережено посилання на джерело – конкретну статтю з корпусу, у якій термін виявлено вперше або найчастіше.

Структура словника передбачає три основні рівні: лексичний (форма терміна та його морфологічні варіанти), семантичний (значення, пов'язані поняття, контекстні відтінки) і статистичний (частота, кількість контекстів, роки появи). У підсумку сформовано 85 ключових термінів, що репрезентують основні напрями сучасних досліджень у галузі штучного інтелекту. Найбільш насиченими виявилися кластери "нейромережеві та генеративні моделі", "нечітка логіка та оптимізація" і "комп'ютерний зір".

Фінальний словник може бути представлений у вигляді таблиці (CSV/Excel), JSON-структури для інтеграції у веб-інтерфейси або традиційного текстового глосарію. Він є завершеним продуктом інформаційної технології лексикографічного аналізу, який демонструє здатність GPT-5 не лише вилучати та узагальнювати терміни, а й створювати цілісну лексикографічну модель наукової предметної галузі.

Фактичний час виконання етапу GPT-5: 7 секунд.

Приклади термінів, включених у словник:

– Згорткова нейронна мережа (CNN) – архітектура для опрацювання даних на ґратках (зображення, спектрограми), де згортки виділяють локальні ознаки; варіанти: *convolutional network*.

– Рекурентна нейронна мережа (RNN) – модель послідовностей із зворотними зв'язками для урахування попередніх станів; варіанти: *LSTM, GRU*.

– Механізм уваги (attention) – вагова схема, що виділяє релевантні частини послідовності/ознакових карт у процесі передбачення.

– Трансформер – архітектура без рекурентності з багатоголовою увагою та позиційним кодуванням; застосування: NLP, зір, мовлення.

– Велика мовна модель (LLM) – трансформер, натренований на масштабних корпусах для генерації/аналізу тексту (узагальнення, класифікація, інструкційні завдання).

– Підлаштування моделей (fine-tuning) – додаткове навчання базової моделі на цільовому домені для підвищення точності.

– Backpropagation – градієнтний алгоритм оновлення ваг НМ шляхом поширення помилки у зворотному напрямі.

– Regularization / Dropout – техніки зменшення перенавчання шляхом обмеження складності або випадкового «вимикання» нейронів.

– GAN (Generative Adversarial Network) – змагальна пара «генератор–дискримінатор» для синтезу даних; проблеми: *mode collapse*, нестабільність.

Нечітка логіка та оптимізація

– Нечіткий висновок – обчислювальна процедура прийняття рішень на базі нечітких правил; етапи: фазифікація, агрегування, дефазифікація.

– Алгоритм Мамдані – класичний підхід нечіткого висновку з використанням мін/макс-операцій і центроїдної дефазифікації.

– Функція належності – відображення [0,1], що задає ступінь належності елемента до нечіткої множини (трикутні, трапецієподібні, ґаусові).

– Багатокритеріальна оптимізація – вибір компромісних рішень за кількома критеріями (час, якість, ресурси); пов'язане: *Парето-оптимум*.

– Нечітка задача комівояжера (fuzzy TSP) – TSP із нечіткими параметрами (вагами), де мета – узгодження критеріїв за функцією згортки.

Комп'ютерний зір, зображення та мовлення

– SIFT / SURF / ORB – інваріантні до масштабу й повороту методи виявлення та опису локальних ознак зображення.

– OpenCV – бібліотека комп'ютерного зору з реалізаціями детекції, трекінгу, калібрування камер, опрацювання зображень.

– ASR (Automatic Speech Recognition) – системи розпізнавання мовлення; приклади: Whisper, Vosk; метрики: WER, RTF.

– Descriptor / Feature Matching – подання локальної ознаки у вигляді вектора та процедура пошуку відповідностей між зображеннями.

Енергетика, смарт-системи та моніторинг

– Розумна енергосистема (Smart Energy) – керування енергоспоживанням і генерацією на базі ШІ для прогнозування й оптимізації навантажень.

– Диференційна модель процесу – рівняння, що описують еволюцію станів системи у часі; застосування: оптимізація, керування.

– Прогнозна аналітика (Predictive Maintenance) – передбачення відмов/зносу обладнання на основі даних сенсорів та моделей RUL/HI.

Освіта, гуманітарні та правові аспекти

– Компетентнісний підхід – навчальна парадигма, орієнтована на вимірювані результати (компетенції) з використанням інструментів ШІ.

– Ситуативне тестування – моделювання навчальних ситуацій та оцінювання рішень за допомогою AI-агентів/LLM.

– Інтелектуальна власність і ШІ – правові режими охорони результатів, створених або згенерованих із участю AI; дискусії щодо авторства й патентоспроможності.

Розподіл отриманих термінів за кластерами наведено в таблиці 1.

Аналіз результатів та оцінка ефективності. Результати проведеного експерименту показали, що застосована інформаційна технологія на базі GPT-5 забезпечує високий рівень автоматизації та якості лексикографічного аналізу. Отримані дані свідчать, що модель здатна виконувати складні багаторівневі операції – від морфологічного розбору до семантичної класифікації – із точністю, порівнянною з роботою фахівця-лінгвіста, але в рази швидше.

У процесі аналізу корпусу GPT-5 продемонструвала високу стабільність частиномовного розпізнавання, що забезпечило чистоту подальших статистичних підрахунків. Середній час на опрацювання одного документа становив менше ніж 0,15 секунди, завдяки чому весь корпус із 51 статті було підготовлено для аналізу практично миттєво. Це підтверджує, що великі мовні моделі можуть ефективно замінювати цілі комплекси традиційних NLP-інструментів, інтегруючи морфологічний, синтаксичний і семантичний рівні опрацювання в єдиній системі.

Ефективність виявлення термінів підтверджується кількісно: із понад 2 400 початкових кандидатів було відібрано 320 термінологічно релевантних одиниць, з яких 200 пройшли нормалізацію й дизамбіґуацію. Такий рівень скорочення (при-

Таблиця 1

Розподіл виділених термінів за кластерами

№	Тематичний кластер	Кількість термінів	Основна характеристика	Приклади ключових термінів
1	Нейромережеві та генеративні моделі	16	Моделі глибинного навчання, трансформери, генеративні мережі, навчання з підкріпленням.	*convolutional neural network, recurrent neural network, transformer, attention mechanism, fine-tuning, GAN, reinforcement learning, backpropagation*
2	Нечітка логіка, оптимізація та еволюційні алгоритми	13	Алгоритмічні підходи до багатокритеріальної оптимізації, нечіткі моделі, евристичні та ройові методи.	*fuzzy logic, membership function, fuzzy TSP, genetic algorithm, swarm intelligence, multi-objective optimization, convolution method*
3	Комп'ютерний зір, мовлення та розпізнавання образів	12	Опрацювання візуальних і звукових даних, методи опису ознак, CNN-архітектури для класифікації зображень.	*feature extraction, image matching, invariant moments, YOLOv3, OpenCV, SIFT, SURF, ORB, ASR, Vosk*
4	Енергетика, промислові та техноекологічні системи	9	Оптимізація енергетичних процесів, моніторинг виробництва, розумні системи управління.	*smart energy, predictive maintenance, topology optimization, additive manufacturing, fuzzy control, renewable resources, process modeling*
5	Багатоагентні системи та кібербезпека	10	Координація агентів, виявлення загроз, діагностика розподілених мереж, графові моделі безпеки.	*multi-agent system, logical diagnostics, cybersecurity model, graph-based optimization, minimax strategy, deep learning fault detection*
6	Освітні, гуманітарні та правові аспекти AI	8	Застосування ШІ в освіті, етичні, правові й соціальні наслідки, компетентнісний підхід.	*competency-based learning, situational testing, academic analytics, AI ethics, intellectual property, authorship of AI-generated works*
7	Онтології, знання та лексикографічні системи	8	Семантичні структури, побудова онтологій, автоматичне вилучення понять і термінів.	*ontology engineering, term extraction, knowledge representation, lexicographic analysis, domain thesaurus*
8	AI у транспорті, агросекторі та UAV-системах	9	Системи навігації безпілотників, моделювання аграрних процесів, контроль трафіку.	*UAV localization, image scene matching, agricultural modeling, yield prediction, situational center, air corridor control*

близно 92%) при збереженні високої точності демонструє відмінну селективну здатність моделі. Фільтрація на основі комбінації метрик C-value, NC-value та Weirdness виявилася оптимальною: показник відсіву надлишкової лексики перевищив 90%, тоді як втрати релевантних одиниць не перевищували 5–7%. Це дозволяє зробити висновок, що GPT-5 здатна стабільно підтримувати баланс між повнотою та чистотою термінологічного корпусу без ручного коригування порогів.

Важливим результатом стало збереження семантичної узгодженості після нормалізації. Попри значне скорочення списку, середній коефіцієнт контекстної подібності між вихідними й уніфікованими формами становив 0.91, тобто більшість термінів зберегли змістову ідентичність. Це свідчить про ефективність механізмів векторної семантики GPT-5, які дають змогу автоматично поєднувати варіанти різних мов, стилів і орфографічних реалізацій без втрати смислу.

Особливу цінність має якість сформованих дефініцій. Середній показник семантичної відповідності визначення до контексту дорівнює 0.88, а експертна оцінка змістової адекватності – 4,2 бала з 5. Це підтверджує, що GPT-5 не лише відтворює формальну структуру дефініції, а й розуміє концептуальну сутність терміна, створюючи визначення, близькі до академічних стандартів. Середній обсяг кожної дефініції – близько 22 слів, що відповідає вимогам до коротких енциклопедичних описів.

Оцінка якості всього процесу показала, що модель підтримує високу стабільність на різних етапах. Узагальнені показники – Precision = 0.86, Recall = 0.78, F1 = 0.82 – підтверджують, що більшість релевантних термінів були виявлені й правильно класифіковані. Навіть за відсутності ручного втручання GPT-5 продемонструвала узгодженість між статистичними, контекстуальними та семантичними критеріями, що є критично важливим для лексикографічних досліджень, орієнтованих на точність дефініцій і зв'язків між поняттями.

Ще одним показником ефективності є структурна когерентність отриманого словника. Аналіз внутрішніх зв'язків між термінами показав середній коефіцієнт подібності 0.73 між спорідненими дефініціями, що означає гармонійність терміносистеми й відсутність суперечностей між близькими поняттями. Це дозволяє розглядати результат не просто як список термінів, а як модель знань, що відображає внутрішню організацію наукової галузі.

Фінальний термінологічний словник із 85 записів сформовано повністю автоматично за менше ніж 10 секунд. Він містить дефініції, частотні показники, контекстні приклади та відомості про рік появи терміна. Загальний час виконання всіх восьми етапів склав 117 секунд, що свідчить про надзвичайну продуктивність GPT-5 порівняно з традиційними інструментами термінографії, де подібний аналіз тривав би кілька годин або днів.

У підсумку можна стверджувати, що GPT-5 забезпечує новий рівень ефективності у сфері лексикографічних досліджень. Вона не лише автоматизує трудомісткі етапи аналізу, а й зберігає наукову достовірність і смислову точність результатів. Модель довела здатність поєднувати швидкість машинного опрацювання з інтерпретативною глибиною, характерною для експертної лексикографії. Отримані результати засвідчують, що великі мовні моделі можуть використовуватись як самостійний інструмент створення цифрових термінологічних баз, де якість не поступається людській, а продуктивність зростає на кілька порядків.

Перспективи подальших досліджень. Результати проведеного дослідження підтвердили, що великі мовні моделі є не просто допоміжним інструментом у лінгвістичних і філологічних студіях, а потенційно – ядром нової парадигми цифрової лексикографії. Їхня здатність інтегрувати синтаксичний, семантичний і прагматичний рівні аналізу текстів робить можливим створення систем, які не лише виявляють терміни, а й розуміють їхні функції, еволюцію та зв'язки у науковому дискурсі. Тому подальші дослідження у цьому напрямі мають бути спрямовані не на заміну людини алгоритмом, а на розроблення інтегрованих людино-машинних систем, здатних до глибокого тлумачення й динамічного оновлення знань.

Першим перспективним напрямом є розбудова масштабних мультимовних і міждисциплінарних корпусів, що дадуть змогу перевірити універсальність LLM у термінографії. Поточні результати, отримані на матеріалі українських наукових текстів у сфері штучного інтелекту, показали ефективність моделі в межах однієї мовної та тематичної системи. Проте подальші експерименти мають охопити порівняльні корпуси різних мов – української, англійської, польської, німецької – для вивчення міжмовних відповідників, концептуальних лакун і варіативності термінів у міжнародному контексті. Це створить умови для побудови універсальних термінологічних онтологій, здатних відображати взаємозв'язки між поняттями незалежно від мовних кордонів.

Ще одним напрямом має стати поглиблення семантичного моделювання. Незважаючи на успіхи сучасних LLM, питання контекстуальної

дизамбігуації, ідентифікації метафоричних уживань і фреймової структури знань залишаються відкритими. Використання гібридних систем, що поєднують нейронні мережі з логічними структурами типу RDF або OWL, може забезпечити більш точну інтерпретацію значень. Такі системи дадуть змогу будувати семантичні графи термінів, у яких кожне поняття матиме визначені відношення типу "є видом", "є частиною", "використовується для", "протистоїть". Цей напрям відкриває можливості для створення глибоких онтологічних карт наукових знань, що стануть базою для освітніх платформ і автоматизованих наукових пошукових систем.

Також перспективним є розроблення динамічних моделей еволюції термінів. Великі мовні моделі мають унікальну властивість – вони можуть навчатися на часових зрізах даних і виявляти зміни в значеннях слів. Це дозволяє створювати хронологічні траєкторії понять, відстежуючи, як термін з'являється, розширює або звужує своє значення, входить у нові дисципліни. У майбутньому це дасть змогу аналізувати темп розвитку науки на рівні лексики – наприклад, виявляти, які концепти з'явилися як нові в певний рік, які втратили актуальність, а які стали міждисциплінарними. Таким чином, лексикографічний аналіз перетворюється на інструмент наукометрії, що дозволяє кількісно оцінювати інтелектуальні тенденції.

Цікавою є можливість використання LLM для побудови інтерактивних і адаптивних термінологічних баз. На відміну від статичних словників, створених вручну, моделі можуть автоматично оновлювати дефініції, доповнювати їх новими контекстами та джерелами. Це дозволить створювати "живі словники", які змінюються разом із розвитком науки. Такі системи можуть функціонувати як відкриті платформи з експертною валідацією, де дослідники доповнюватимуть або коригуватимуть результати моделі. У перспективі це може привести до формування колаборативних термінологічних екосистем, де людина і штучний інтелект працюватимуть спільно у процесі творення наукового знання.

Крім цього, автоматичне виділення термінів і створення дефініцій дає змогу формувати тематичні лексикони для студентів, розробляти адаптивні навчальні матеріали, оцінювати рівень термінологічної компетентності. У майбутньому можливою стане поява інтелектуальних навчальних систем, які в реальному часі аналізуватимуть тексти студентів, пропонуватимуть корекцію термінів або пояснення їхніх значень. Це дозволить інтегрувати результати лексикографічних досліджень у практику освіти й підготовки науковців.

Таким чином, перспективи подальших досліджень полягають у переосмисленні самої природи термінографії – від статичного опису термінів до створення самонавчальних систем знань. Великі мовні моделі поступово стають не лише технічним інструментом, а співдослідником, який бере участь у формуванні наукових категорій, відображаючи їхній розвиток, взаємозв'язки та культурний контекст. Саме в цьому полягає стратегічна мета наступного етапу – побудова інтелектуальної екосистеми, де наука, мова й технології зливаються в єдину пізнавальну структуру.

Висновки. Проведене дослідження довело, що запропонована інформаційна технологія лексикографічного аналізу на основі великих мовних моделей забезпечує повний, керований і відтворюваний перехід від сирого корпусу наукових текстів до компактного, структурованого термінологічного словника. Конвеєр із восьми кроків – від підготовки корпусу до формування вихідного словника – виявився узгодженим як методологічно, так і обчислювально. Ключова ідея – поєднати лінгвістичні евристики (шаблони частин мови, робота з абревіатурами), статистичні критерії «термінності» (C-value/NC-value, індекс спеціалізації), графові підходи та семантичні подання – дала змогу досягти балансу між повнотою й точністю без потреби в тривалому ручному налаштуванні.

Емпірична частина підтвердила придатність технології для реального академічного контексту журналу «Штучний інтелект»: ми опрацювали 51 статтю за останні роки та побудували на їх основі короткий галузевий міні-словник. Вибір такого джерела й масштаб корпусу забезпечили тематичну однорідність, але водночас – достатню різноманітність жанрів та підтем, що важливо для перевірки стійкості алгоритмів до стилістичних зрушень і міждисциплінарних перетинів. Формально постановку задачі експерименту – отримати компактний словник термінів ШІ із дефініціями – було закріплено в описі дослідницького дизайну й складу корпусу (51 публікація) та узгоджено з метою побудови практично корисного результату для термінографії й освіти.

З технічного погляду найбільший приріст якості дала комбінація фільтрації кандидатів за статистичними метриками та подальша семантична перевірка через векторні подання. Це дозволило різко зменшити початковий список кандидатів, зберігши концептуальне ядро предметної області. При цьому уніфікація варіантів – завдяки зіставленню абревіатур і повних форм, а також кластеризації близьких лексем – зняла проблему дублювання і підвищила когерентність терміно-

системи. В результаті ми отримали компактний набір уніфікованих термінів і змогли додати для них короткі дефініції, що семантично узгоджуються з корпусними контекстами. Узгодження зі змістом статті, де акцентують потребу долати статичність і фрагментарність традиційної термінографії, підтверджує практичний сенс автоматизації саме в запропонованій формі.

Змістові підсумки словника свідчать, що найбільш «щільними» виявилися напрями сучасного ШІ, очікувані для останніх років: нейромережеві й генеративні моделі, нечітка логіка та оптимізація, комп'ютерний зір, онтології знань і багатоагентні системи. Це добре корелює з розміткою кластерів результатів і переліком характерних термінів, засвідчуючи, що автоматична технологія здатна віддзеркалювати актуальну структуру наукової галузі.

Окремо відзначимо, що інструментальною перевагою LLM стало контекстне розмежування значень: автоматичне групування вживань термінів у типових для журналу тематичних контекстах (наприклад, «модель» як архітектура нейромережі проти «моделі» як математичного опису процесу) зменшило кількість семантичних помилок у дефініціях і знизило ризик омонімії. Це узгоджується з концепцією кроку про дизамбіґуацію, де визначено потребу розводити значення за контекстними кластерами, перш ніж фіксувати дефініції.

Отримані кількісні показники якості – збалансоване співвідношення точності й повноти, висока семантична узгодженість дефініцій із корпусом – підтверджують працездатність підходу для україномовного наукового дискурсу зі значною часткою англомовних запозичень. На практиці це означає можливість регулярного (періодичного) оновлення словника без повного перезапуску процесу: конвеєр дозволяє дозавантажувати нові статті, ідентифікувати нові або еволюціонуючі терміни та інтегрувати їх у вже наявну структуру. З погляду прикладних доменів журналу це відкриває шлях до побудови «живих» глосаріїв для навчання, бібліотечних сервісів і дослідницьких інформаційних систем.

Обмеження дослідження пов'язані з профілем корпусу та притаманними великим мовним моделям ризиками. По-перше, вибірка з одного фахового видання забезпечує чистоту термінів, але зменшує жанрову різноманітність; розширення корпусу за рахунок суміжних журналів і матеріалів конференцій може підвищити стійкість моделі до стилістичних варіацій і неочевидних термінотворень. По-друге, попри успішну нормалізацію, багатомовна варіативність (українські та англійські форми, локальні транслітерації) залишає «хвости» рідкісних написань, які потребуватимуть додаткового зворотного зв'язку користувачів. По-третє, повністю автоматичне формування дефініцій у складних випадках (міждисциплінарні перетини, нові метафоричні вживання) все ще виграє від короткої експертної валідації – зокрема для стандартизації стилю термінів у навчальних матеріалах і довідниках.

Водночас навіть за цих обмежень технологія вже сьогодні розв'язує головні «вузькі місця» традиційної термінографії: швидкість оновлення, цілісність подання, уніфікацію варіантів і зв'язність між поняттями. З методологічного погляду запропонований підхід дає три важливі ефекти. Перший – операційний: повний цикл від завантаження статей до готового словника вкладається у хвилини, що робить можливим регулярне оновлення терміносистеми разом із публікаційним потоком. Другий – якісний: поєднання статистики й семантики зменшує шум і підвищує інтерпретованість, тож результат придатний не лише для пошуку, а й для навчання та наукового огляду. Третій – масштабований: архітектура конвеєра дозволяє переносити рішення на інші предметні області (наприклад, біомедицину чи правові дослідження) без перегляду базових етапів; змінюються передусім ресурси (корпус, словники-референти) і пороги фільтрації.

Підсумовуючи, робота продемонструвала, що великі мовні моделі є практично корисним ядром цифрової термінографії. Вони не заміняють фахівця, але беруть на себе рутину – витяг, нормалізацію, узагальнення – і постачають узгоджений матеріал для експертної валідації та подальшої стандартизації. На прикладі корпусу статей «Штучного інтелекту» ми отримали компактний, змістовно збалансований словник, який відображає реальну структуру сучасних досліджень і готовий до інтеграції в освітні, бібліотечні та науково-аналітичні системи. З огляду на зафіксовані показники та чітко окреслені межі застосування, технологія може стати основою «живих» терміносистем для української наукової спільноти – з регулярним оновленням, прозорою методологією й мінімальною ціною підтримки.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Xu K., Feng Y., Li Q., Dong Z., Wei J. Survey on terminology extraction from texts. Journal of Big Data. 2025. Vol. 12. №. 29. DOI:10.1186/s40537-025-01077-x.

2. Shao W., Hua B., Song L. A Pattern and POS Auto-Learning Method for Terminology Extraction from Scientific Texts. Data and Information Management. 2021. Vol. 5. №. 3. P. 329-342. DOI:10.1016/j.dim.2021.100003.

3. Wissik T. Impact of automatic term extraction on terminology work. Terminology. 2025. Vol. 31. №. 1. P. 110-135. DOI:10.1075/term.00085.wis.

4. Martinez-Rodriguez J.L., Hogan A., Lopez-Arevalo I. Information Extraction meets the Semantic Web: A Survey. Semantic Web. 2018. Vol. 11. №. 2. P. 1-81. DOI:10.3233/SW-180333.

5. Pazienza M.T., Pennacchiotti M., Zanzotto F.M. Terminology Extraction: An Analysis of Linguistic and Statistical Approaches. Knowledge Mining. Springer, 2008. P. 25-50. DOI:10.1007/978-3-540-77803-3_3.

6. Frantzi K., Ananiadou S., Mima H. Automatic Recognition of Multi-word Terms: The C-value/NC-value Method. International Journal on Digital Libraries. 2000. Vol. 3. №. 2. P. 115-130. DOI:10.1007/s007999900023.

7. Mihalcea R., Tarau P. TextRank: Bringing Order into Text. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. 2004. P. 404-411. URL: https://aclanthology.org/W04-3252/.

8. Hearst M.A. Automatic Acquisition of Hyponyms from Large Text Corpora. Proceedings of the 14th Conference on Computational Linguistics. 1992. Vol. 2. P. 539-545. DOI:10.3115/992133.992154.

9. Schwartz A.S., Hearst M.A. A Simple Algorithm for Identifying Abbreviation Definitions in Biomedical Text. Proceedings of the Pacific Symposium on Biocomputing. 2003. P. 451-462.

10. Dagdelen J. et al. Structured Information Extraction from Scientific Text with Large Language Models. Nature Communications. 2024. Vol. 15. №. 1418. DOI:10.1038/s41467-024-45563-x.

11. Chatzitheodorou A. et al. KnowledgeTB: Leveraging Large Language Models for Enhanced Terminology Extraction over Knowledge Graphs. CEUR Workshop Proceedings. 2025. Vol. 3990. URL: https://ceur-ws.org/Vol-3990/.

12. Tran H.T.H., Martinc M. et al. The Recent Advances in Automatic Term Extraction: A Survey. arXiv:2301.06767. 2023. URL: https://doi.org/10.48550/arXiv.2301.06767.

13. Minaee K. et al. Large Language Models: A Survey. arXiv:2402.06196. 2024. URL: https://doi.org/10.48550/arXiv.2402.06196.

14. Zhao W.X. et al. A Survey of Large Language Models. arXiv:2303.18223. 2023. URL: https://doi.org/10.48550/arXiv.2303.18223.

15. Chun Y. et al. Enhancing Automatic Term Extraction with Large Language Models via Syntactic Retrieval. Findings of the Association for Computational Linguistics: ACL. 2025. P. 9916-9926. DOI:10.18653/v1/2025.findings-acl.516.

16. Liu Y., Song Y., Liu H. Adaptive POS Embedding for Scientific Term Extraction. Computational Linguistics. 2024.

17. Shcheglova M., Korotkova N., Zhiltsova E. Knowledge Graph-Enhanced Term Extraction with Large Language Models. Information Systems Frontiers. 2025.

18. De Lara P., Martins B., Costa M. Cross-lingual Terminology Extraction via Semantic Alignment. Expert Systems with Applications. 2023.

19. Moreira-Filho J.T. et al. Automating Data Extraction from Scientific Literature Using Large Language Models. Wiley Interdisciplinary Reviews: Computational Molecular Science. 2025. Vol. 15. №. 5. Art. e70047. DOI:10.1002/wcms.70047.

20. Wang R., Yao T., Zhou J. Terminology-BERT: Domain-Adaptive Pre-training for Term Extraction. Knowledge-Based Systems. 2024.

21. TExEval-2024 Shared Task Results. Proceedings of the European Language Resources Association (ELRA). 2024.

22. Mamontova A., Ischenko R., Vorontsov K. RuTermEval-2024: Cross-domain Term Extraction and Classification. Dialogue Conference Proceedings. 2025. P. 251-256. DOI:10.28995/2075-7182-2025-21-2(51)-251-256.

23. Lyu H., Guo C., Yang S. Automatic Glossary Generation from Academic Texts Using LLMs. Digital Scholarship in the Humanities. 2025.

24. Hassan A., Mahmoud M., El-Baz D. Ontology Construction in Biomedicine Using Large Language Models. Artificial Intelligence in Medicine. 2024.

25. Zhang X., Li H., Wang D. From Keywords to Knowledge Units: A Conceptual Approach to Automatic Term Extraction. Information Processing & Management. 2025.

26. Пасічник В., Яромич М. Великі мовні моделі та онтології у філологічних дослідженнях: аналітичний огляд джерел. Актуальні питання гуманітарних наук. Мовознавство. Літературознавство. 2025. Вип. 83. №. 3. С. 236–250. DOI:10.24919/2308-4863/83-3-35.

27. Кунанець Н., Яромич М. Виділення концептів у літературних текстах із використанням великих мовних моделей. Вісник науки та освіти. 2025. Вип. 32. №. 2. С. 343–357. DOI:10.52058/2786-6165-2025-2(32)-343-357.

28. Пасічник В., Яромич М. Автоматизоване формування технічної документації в галузі ІТ з використанням великих мовних моделей. Studia Methodologica. 2025. Вип. 59. С. 250-273. DOI:10.32782/2307-1222.2025-59-22.

29. Пасічник В., Яромич М. Порівняльний аналіз великих мовних моделей в контексті вирішення лінгвістичних задач. Актуальні питання гуманітарних наук. Мовознавство. Літературознавство. 2025. Вип. 89. №. 1. С. 288–305. DOI:10.24919/2308-4863/89-1-41.

30. Пасічник В., Яромич М. Особливості жанрової класифікації літератури за допомогою великих мовних моделей. Folium. 2025. №. 6. С. 132–143. DOI:10.32782/folium/2025.6.19.

31. Пасічник В., Яромич М. Методи та засоби великих мовних моделей для концептуалізації предметної області "штучний інтелект". Технічні науки : тези доп. IX Міжнародної науково-практичної конференції "Science In The Modern World: Innovations And Challenges", 15-17.05.2025, Торонто, Канада. 2025. С. 353–358.

REFERENCES

1. Xu, K., Feng, Y., Li, Q., Dong, Z., & Wei, J. (2025). Survey on terminology extraction from texts. Journal of Big Data. Vol. 12. №. 29. https://doi.org/10.1186/s40537-025-01077-x

2. Shao, W., Hua, B., & Song, L. (2021). A Pattern and POS Auto-Learning Method for Terminology Extraction from Scientific Texts. Data and Information Management. Vol. 5. №. 3. P. 329–342. https://doi.org/10.1016/j.dim.2021.100003

3. Wissik, T. (2025). Impact of automatic term extraction on terminology work. Terminology. Vol. 31. №. 1. P. 110–135. https://doi.org/10.1075/term.00085.wis

4. Martinez-Rodriguez, J. L., Hogan, A., & Lopez-Arevalo, I. (2018). Information Extraction meets the Semantic Web: A Survey. Semantic Web. Vol. 11. №. 2. P. 1–81. https://doi.org/10.3233/SW-180333

5. Pazienza, M. T., Pennacchiotti, M., & Zanzotto, F. M. (2008). Terminology Extraction: An Analysis of Linguistic and Statistical Approaches. Knowledge Mining. P. 25–50. https://doi.org/10.1007/978-3-540-77803-3_3

6. Frantzi, K., Ananiadou, S., & Mima, H. (2000). Automatic Recognition of Multi-word Terms: The C-value/NC-value Method. International Journal on Digital Libraries. Vol. 3. №. 2. P. 115–130. https://doi.org/10.1007/s007999900023

7. Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing Order into Text. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. P. 404–411. Retrieved from: https://aclanthology.org/W04-3252/

8. Hearst, M. A. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. Proceedings of the 14th Conference on Computational Linguistics. Vol. 2. P. 539–545. https://doi.org/10.3115/992133.992154

9. Schwartz, A. S., & Hearst, M. A. (2003). A Simple Algorithm for Identifying Abbreviation Definitions in Biomedical Text. Proceedings of the Pacific Symposium on Biocomputing. P. 451–462.

10. Dagdelen, J. et al. (2024). Structured Information Extraction from Scientific Text with Large Language Models. Nature Communications. Vol. 15. №. 1418. https://doi.org/10.1038/s41467-024-45563-x

11. Chatzitheodorou, A. et al. (2025). KnowledgeTB: Leveraging Large Language Models for Enhanced Terminology Extraction over Knowledge Graphs. CEUR Workshop Proceedings. Vol. 3990. Retrieved from: https://ceur-ws.org/Vol-3990/

12. Tran, H. T. H., Martinc, M. et al. (2023). The Recent Advances in Automatic Term Extraction: A Survey. arXiv:2301.06767. Retrieved from: https://doi.org/10.48550/arXiv.2301.06767

13. Minaee, K. et al. (2024). Large Language Models: A Survey. arXiv:2402.06196. Retrieved from: https://doi.org/10.48550/arXiv.2402.06196

14. Zhao, W. X. et al. (2023). A Survey of Large Language Models. arXiv:2303.18223. Retrieved from: https://doi.org/10.48550/arXiv.2303.18223

15. Chun, Y. et al. (2025). Enhancing Automatic Term Extraction with Large Language Models via Syntactic Retrieval. Findings of the Association for Computational Linguistics: ACL. P. 9916–9926. https://doi.org/10.18653/v1/2025.findings-acl.516

16. Liu, Y., Song, Y., & Liu, H. (2024). Adaptive POS Embedding for Scientific Term Extraction. Computational Linguistics.

17. Shcheglova, M., Korotkova, N., & Zhiltsova, E. (2025). Knowledge Graph-Enhanced Term Extraction with Large Language Models. Information Systems Frontiers.

18. De Lara, P., Martins, B., & Costa, M. (2023). Cross-lingual Terminology Extraction via Semantic Alignment. Expert Systems with Applications.

19. Moreira-Filho, J. T. et al. (2025). Automating Data Extraction from Scientific Literature Using Large Language Models. Wiley Interdisciplinary Reviews: Computational Molecular Science. Vol. 15. №. 5. Art. e70047. https://doi.org/10.1002/wcms.70047

20. Wang, R., Yao, T., & Zhou, J. (2024). Terminology-BERT: Domain-Adaptive Pre-training for Term Extraction. Knowledge-Based Systems.

21. TExEval-2024 Shared Task Results. (2024). Proceedings of the European Language Resources Association (ELRA).

22. Mamontova, A., Ischenko, R., & Vorontsov, K. (2025). RuTermEval-2024: Cross-domain Term Extraction and Classification. Dialogue Conference Proceedings. P. 251–256. https://doi.org/10.28995/2075-7182-2025-21-2(51)-251-256

23. Lyu, H., Guo, C., & Yang, S. (2025). Automatic Glossary Generation from Academic Texts Using LLMs. Digital Scholarship in the Humanities.

24. Hassan, A., Mahmoud, M., & El-Baz, D. (2024). Ontology Construction in Biomedicine Using Large Language Models. Artificial Intelligence in Medicine.

25. Zhang, X., Li, H., & Wang, D. (2025). From Keywords to Knowledge Units: A Conceptual Approach to Automatic Term Extraction. Information Processing & Management.

26. Pasichnyk, V. V., & Yaromych, M. V. (2025). Velyki movni modeli ta ontolohii u filolohichnykh doslidzhenniakh: analitychnyi ohliad dzherel [Large language models and ontologies in philological research: An analytical review of sources]. Aktualni pytannia humanitarnykh nauk. Movoznavstvo. Literaturoznavstvo. Vol. 83. №. 3. P. 236–250. https://doi.org/10.24919/2308-4863/83-3-35. [in Ukrainian].

27. Kunanets, N. E., & Yaromych, M. V. (2025). Vydilennia kontseptiv u literaturnykh tekstakh iz vykorystanniam velykykh movnykh modelei [Extracting concepts from literary texts using large language models]. Visnyk nauky ta osvity. Vol. 32. №. 2. P. 343–357. https://doi.org/10.52058/2786-6165-2025-2(32)-343-357. [in Ukrainian].

28. Pasichnyk, V. V., & Yaromych, M. V. (2025). Avtomatyzovane formuvannia tekhnichnoi dokumentatsii v haluzi IT z vykorystanniam velykykh movnykh modelei [Automated generation of technical documentation in the IT field using large language models]. Studia Methodologica. Vol. 59. P. 250–273. https://doi.org/10.32782/2307-1222.2025-59-22. [in Ukrainian].

29. Pasichnyk, V. V., & Yaromych, M. V. (2025). Porivnialnyi analiz velykykh movnykh modelei v konteksti vyrishennia linhvistychnykh zadach [Comparative analysis of large language models in the context of solving linguistic tasks]. Aktualni pytannia humanitarnykh nauk. Movoznavstvo. Literaturoznavstvo. Vol. 89. №. 1. P. 288–305. https://doi.org/10.24919/2308-4863/89-1-41. [in Ukrainian].

30. Pasichnyk, V. V., & Yaromych, M. V. (2025). Osoblyvosti zhanrovoi klasyfikatsii literatury za dopomohoiu velykykh movnykh modelei [Features of genre classification of literature using large language models]. Folium. Vol. 6. P. 132–143. https://doi.org/10.32782/folium/2025.6.19. [in Ukrainian].

31. Pasichnyk, V. V., & Yaromych, M. V. (2025). Metody ta zasoby velykykh movnykh modelei dlia kontseptualizatsii predmetnoi oblasti "shtuchnyi intelekt" [Methods and tools of large language models for conceptualization of the subject area "artificial intelligence"]. Tekhnichni nauky: tezy dop. IX Mizhnarodnoi naukovo-praktychnoi konferentsii "Science in the Modern World: Innovations and Challenges", 15–17.05.2025, Toronto, Kanada. P. 353–358 [in Ukrainian].

Дата першого надходження рукопису до видання: 26.10.2025
Дата прийнятого до друку рукопису після рецензування: 28.11.2025
Дата публікації: 19.12.2025