КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 2 (34), 2026
ISSN 2663 – 4023
DOI 10.28925/2663-4023.2026.34.1226
УДК 004.56:004.85

Винар Андрій Сергійович
магістр з напряму Інженерія програмного забезпечення
Вищий навчальний заклад «Міжрегіональна Академія управління персоналом», Київ, Україна
ORCID: 0000-0003-2044-7590
andrii.vynar.om@gmail.com

Сведенюк Олексій Валерійович
аспірант кафедри захисту інформації
Національний Університет «Львівська Політехніка», Львів, Україна
ORCID: 0009-0001-6125-7382
oleksii.svedeniuk.asp.2025@lpnu.ua

Мединський Остап Романович
аспірант кафедри захисту інформації
Національний Університет «Львівська Політехніка», Львів, Україна
ORCID: 0009-0003-6582-828X
ostap.medynskyi.asp.2025@lpnu.ua

ОЦІНЮВАННЯ ЕФЕКТИВНОСТІ RAG-АРХІТЕКТУР ДЛЯ АВТОМАТИЗАЦІЇ АНАЛІЗУ ПОЛІТИК ІНФОРМАЦІЙНОЇ БЕЗПЕКИ

Анотація. Стаття присвячена емпіричному порівнянню ефективності трьох ключових архітектур Retrieval-Augmented Generation (RAG) – базової (Naive RAG), покращеної (Advanced RAG) та агентно-орієнтованої (Agentic RAG). Метою дослідження є аналіз їх ефективності у генерації точних і повних відповідей на запитання в критичній сфері інформаційної безпеки на основі внутрішніх корпоративних політик та процедур. Для проведення експериментів було застосовано спеціалізований Python-фреймворк із бібліотекою LangChain, векторною базою даних ChromaDB, а також гібридну конфігурацію локальних (Ollama: LLaMA 3) та хмарної (OpenAI ChatGPT-5) LLM-платформ. Результати демонструють значний потенціал застосування Advanced та Agentic RAG для підвищення точності, повноти і контекстуальної релевантності відповідей. Запропонований підхід підтверджує свою ефективність для автоматизації та спрощення процесу опрацювання великого обсягу документів фахівцями з інформаційної безпеки та аудиторами.

Ключові слова: Retrieval-Augmented Generation, Agentic RAG, Advanced RAG, LangChain, ChatGPT-5, інформаційна безпека, LLM.

ВСТУП

Стрімкий розвиток великих мовних моделей (LLM) відкриває нові можливості для автоматизації інтелектуальних задач у різних галузях, зокрема в інформаційній безпеці. Одним із найперспективніших напрямів є застосування технології Retrieval-Augmented Generation (RAG) — підходу, що поєднує семантичний пошук у корпоративних документах із генеративними можливостями сучасних мовних моделей. На відміну від традиційних LLM, обмежених статичними знаннями часу навчання, RAG-системи здатні динамічно звертатись до актуальних внутрішніх джерел, формуючи обґрунтовані та контекстуально релевантні відповіді на основі реальної нормативної бази організації.

Особливої актуальності це набуває в контексті аудиту інформаційної безпеки, де фахівці щоденно опрацьовують значні масиви внутрішніх політик, процедур та інструкцій відповідно до вимог стандартів (зокрема ISO/IEC 27001). Традиційний ручний підхід до підготовки відповідей на аудиторські запити є трудомістким, схильним до суб'єктивних неточностей та погано масштабованим при зростанні обсягу документації. Автоматизація цього процесу за допомогою RAG-систем може суттєво підвищити ефективність роботи служб інформаційної безпеки та аудиторських організацій.

Проте сучасний ландшафт RAG-рішень є неоднорідним: від простих лінійних архітектур (Naive RAG) до складних агентно-орієнтованих підходів (Agentic RAG) із механізмами самокорекції та

ітеративного пошуку. Кожна архітектура демонструє різну ефективність залежно від складності запитів і характеру оброблюваних документів. Незважаючи на активні дослідження у цій сфері, практичне емпіричне порівняння різних RAG-архітектур саме в задачах інформаційної безпеки залишається недостатньо вивченим.

У зв'язку з цим, дана стаття присвячена порівняльному аналізу трьох ключових архітектур RAG — базової (Naive RAG), покращеної (Advanced RAG) та агентно-орієнтованої (Agentic RAG) — у контексті генерації відповідей на запитання з інформаційної безпеки на основі внутрішньої корпоративної документації. Дослідження спирається на реальний тестовий набір з 50 політик і 1000 запитань, а оцінювання здійснюється за метриками контекстуальної релевантності, повноти та фактологічної точності з використанням двох LLM-платформ: локальної (LLaMA 3.2 via Ollama) та хмарної (OpenAI ChatGPT-5).

Постановка проблеми. Сучасні процеси аудиту інформаційної безпеки у великих організаціях критично залежать від своєчасної роботи і доступу до внутрішньої нормативної документації – політик, процедур та інструкцій. Традиційно підготовка відповідей на запити аудиторів чи клієнтів здійснюється вручну фахівцем з інформаційної безпеки, що є трудомістким процесом і часто призводить до суб'єктивних неточностей та розбіжностей в інтерпретації. Аналіз великих масивів корпоративних документів потребує надійної автоматизації. Водночас внутрішні політики та процедури інформаційної безпеки базуються на загальновизнаних стандартах та вимогах [5], тому виникає потреба у системі, здатній швидко знаходити в цих документах відповідні положення.

Одним з перспективних підходів до вирішення цієї проблеми є Retrieval-Augmented Generation (RAG) – поєднання механізмів пошуку інформації та генерації тексту великими мовними моделями. RAG дозволяє долати обмеження статичних знань моделі, отримуючи актуальний контекст з автентичних внутрішніх джерел перед формуванням відповіді. Таким чином, модель RAG може формувати більш точні та обґрунтовані відповіді з урахуванням специфічних внутрішніх політик компанії, підвищуючи достовірність і практичну цінність результатів [13].

Аналіз останніх досліджень і публікацій. В останні роки спостерігається швидка еволюція архітектур RAG відповідно до зростаючих вимог точності та складності запитів [4]. Базова архітектура (Naive RAG), яка візуально зображена на Рис. 1, реалізує лінійну схему «запит → пошук → генерація», яка є простою та швидкою у впровадженні, проте чутливою до якості пошукового запиту і може повертати неповні чи нерелевантні відповіді при обмеженому обсязі вибірки контексту.

Рис.1 Базова архітектура (Naive RAG).

Зображена на Рис. 2, покращена архітектура (Advanced RAG) включає додаткові механізми оптимізації пошуку – зокрема, пошук з використанням декількох запитів (Multi-Query Retriever) для підвищення повноти контексту та переосмислення знайдених фрагментів за допомогою додаткових моделей або метрик релевантності [10]. Також можуть комбінуватися різні підходи до пошуку (dense + sparse retrieval) для охоплення максимально релевантної інформації [7].

Рис 2. Покращена архітектура (Advanced RAG)

Агентно-орієнтована архітектура (Agentic RAG, також Modular RAG), яка зображена на Рис. 3, додатково впроваджує автономні програмні агенти, які здатні динамічно керувати стратегією пошуку та генерації.

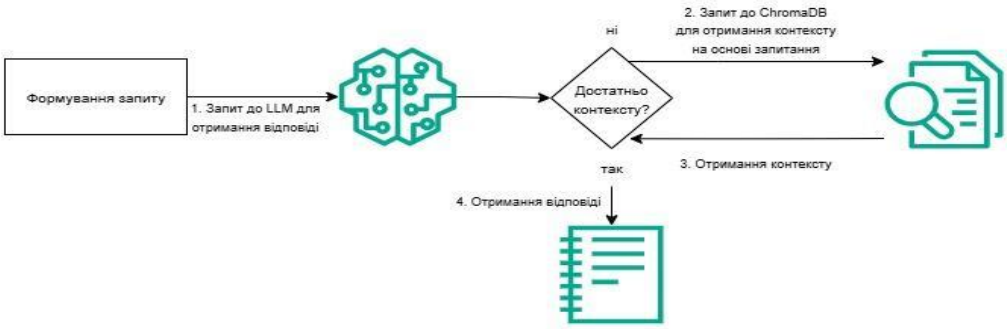

Рис 3. Агентно-орієнтована архітектура (Agentic RAG)

Такі агенти можуть інтерактивно уточнювати запит, виконувати додаткові пошукові під задачі та використовувати інструменти для обробки даних, реалізуючи тим самим багатокроковий пошук з само корекцією помилок [12]. Вбудовування агентів у RAG-пайплайн підвищує гнучкість і контекстну обізнаність системи, дозволяючи ефективно відповідати навіть на складні багатоетапні запити [11].

У своїй статті “Методи реалізації retrieval-augmented generation у поєднанні з сучасними великими мовними моделями” Олійник і Чичкарьов проаналізували сучасні методи RAG у поєднанні з моделями GPT-4-рівня та продемонстрували, що інноваційні підходи (спільне навчання компонентів, адаптивний динамічний пошук, self-refine із зворотним зв’язком, модульна інтеграція моделей через API) дозволяють долати недоліки традиційного RAG – знижувати рівень «галюцинацій» LLM і підвищувати фактичну точність відповідей [3]. Також варто зазначити, що застосування методик RAG, Advanced RAG та Agentic RAG з фреймворком MITRE ATT&CK дозволяє автоматизувати моделювання загроз і розробку стратегій кіберзахисту шляхом інтелектуального вилучення, структурування та контекстуального аналізу великої бази даних тактик і технік реальних зловмисників [1]. Це підтверджує високий потенціал вдосконалених RAG-підходів для покращення якості генерованих відповідей у різноманітних галузях, включно з інформаційною безпекою для побудови комплексної моделі безпеки організації [2].

Незважаючи на активний розвиток методів RAG у загальному випадку [4, 13], їх практичне порівняння саме в галузі інформаційної безпеки проводилось досить рідко. Це робить дане дослідження актуальним і затребуваним як для компаній, що проходять аудити інформаційної безпеки, так і для аудиторських організацій, які прагнуть автоматизувати аналіз відповідності внутрішніх політик стандартам. Кожна з архітектур RAG має свої переваги та недоліки, що проявляються залежно від сценарію використання. Очікувано, більш просунуті підходи (Advanced і Agentic RAG) повинні забезпечити вищу якість відповідей у контексті наших задач. На основі проведеного аналізу поставлено гіпотезу, що саме Advanced RAG і Agentic RAG найкраще підходять для надання високоякісних відповідей на питання інформаційної безпеки, оскільки поєднують ширше охоплення контексту та механізми самокорекції. Це твердження буде перевірено експериментально в межах даної роботи.

Мета статті. Мета даної статті полягає в аналізі та порівнянні ефективності застосування різних архітектур Retrieval-Augmented Generation (RAG) для оптимізації процесів інформаційної та кібербезпеки, зокрема для швидкого та точного надання відповідей на аудиторські опитувальники на основі внутрішніх нормативних документів компанії. Реалізація цієї мети передбачає експериментальне оцінювання якості відповідей, згенерованих базовою, покращеною та агентною RAG-моделями, та визначення переваг і обмежень кожної архітектури в контексті задач інформаційної безпеки.

Методологія дослідження. Для досягнення поставленої мети було розгорнуто експериментальну систему, яка поєднує модуль пошуку і генеративну модель. В ролі сховища знань були використані внутрішні документи компанії у сфері інформаційної безпеки (політики, процедури, інструкції у форматах PDF, DOCX, TXT). Вони були конвертовані у векторні представлення та завантажені до векторної бази ChromaDB. Документи розбито на фрагменти обсягом ~700 токенів із перекриттям ~150 токенів для збереження контексту при пошуку. Для побудови векторного індексу використано модель ембеддингу all-MiniLM-L6-v2, яка забезпечує збалансовану точність та швидкодію при пошуку [6]. Використані LLM-платформи.

Генерація відповідей здійснювалася двома типами великих мовних моделей:

OpenAI ChatGPT-5 – хмарна LLM останнього покоління, доступ до якої здійснювався через API [8]; локальні моделі на базі LLaMA 3.2, розгорнуті через платформу Ollama на сервері для експериментів без передачі конфіденційних даних у хмару.

Архітектурна реалізація RAG. Зв'язок між модулем пошуку та LLM реалізовано за допомогою фреймворку LangChain, який надає уніфікований інтерфейс для побудови різних ланцюжків RAG [9]. Це дозволило з відносною легкістю переключатися між реалізаціями Naive, Advanced (зокрема з Multi-Query Retriever) та Agentic RAG для їхнього подальшого тестування.

Оцінювання та метрики. Для об'єктивної оцінки ефективності підходів сформовано тестовий набір, що включає 50 власноруч підготовлених політик інформаційної безпеки та 1000 запитань, які охоплюють ключові напрями (керування доступом, управління інцидентами, резервне копіювання, навчання персоналу, тощо). Якість згенерованих відповідей оцінювалася за трьома ключовими метриками: контекстуальна релевантність (наскільки відповідь стосується саме того контексту політик, що запитується), повнота відповіді (чи охоплено всі аспекти питання) та фактологічна точність (відсутність фактичних помилок, відповідність інформації політикам).

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

В рамках експерименту порівнювалися три, звичайна Naive RAG та вдосконалені RAG-архітектури – Advanced RAG з механізмом Multi-Query і Modular/Agentic RAG з використанням агентів-інструментів – на двох платформах LLM: локальній моделі LLaMA 3.2 та хмарній моделі ChatGPT-5. Таблиця 1 наводить середні показники якості відповідей згідно з описаними метриками для кожної комбінації архітектури та моделі.

Таблиця 1

Порівняльні показники контекстуальної релевантності, повноти та фактологічної точності відповідей для Advanced RAG (multi-query) і Agentic RAG (tool calling) на моделях LLaMA 3.2 та ChatGPT-5.

Архітектура RAG (LLM)	Контекстуальна релевантність	Повнота відповіді	Фактологічна точність
Naïve RAG (LLaMA 3.2)	0.49	0.5	0.52
Naïve RAG (ChatGPT-5)	0.49	0.52	0.55
Advanced RAG (LLaMA 3.2)	0,86	0,78	0,89
Advanced RAG (ChatGPT-5)	0,86	0,85	0,95
Agentic RAG (LLaMA 3.2)	0,95	0,89	0,90
Agentic RAG (ChatGPT-5)	0,95	0,94	0,96

Як видно з таблиці, базова архітектура (Naive RAG) суттєво поступається покращеним підходам. За всіма показниками якості Modular/Agentic RAG продемонструвала найвищі результати. Якість відповідей за всіма метриками перевищує Naive RAG на 35–40%, що підтверджує ефективність агентно-орієнтованої логіки в даній задачі. Це означає, що агентний підхід краще охоплює всі аспекти запитання завдяки можливості виконувати додаткові ітерації пошуку та використовувати допоміжні інструменти. Варто відзначити, що різниця між Advanced та Agentic RAG менш суттєва – обидві архітектури показують високий рівень точності та повноти, перевершуючи базовий підхід. Отже, використання додаткових механізмів таких як, Multi-Query, переписування запиту, ітеративний пошук тощо, значно покращує результати генерації у сфері інформаційної безпеки.

Варте уваги також те, що різниця між локальною та хмарною LLM проявляється у показниках повноти та точності: модель ChatGPT-5 загалом перевершує LLaMA 3.2 у межах кожної архітектури (наприклад, фактологічна точність Advanced RAG зросла з 0,89 до 0,95 при переході від LLaMA до ChatGPT). Це очікувано, адже більш потужна модель краще інтерпретує знайдений контекст і формулює відповіді. Водночас порівняння між архітектурами показує, що навіть з простішою мовною моделлю (LLaMA 3.2) агентний підхід дає змогу наблизитись за якістю до результатів потужнішої моделі в

покращеній архітектурі (наприклад, повнота 0,89 у Agentic RAG на LLaMA проти 0,85 у Advanced RAG на ChatGPT).

Отримані результати підтверджують, що Advanced RAG як вдосконалення базового підходу суттєво підвищує повноту та точність відповідей завдяки використанню мульти-пошукових запитів і оптимізації вибірки контексту. Водночас Agentic RAG забезпечує ще вищу якість за рахунок динамічної адаптації. Агенти можуть ідентифікувати, коли перший отриманий контекст є недостатнім, і автоматично виконувати додаткові пошукові ітерації чи застосовувати інструменти (наприклад, для перефразування запиту або аналізу знайдених даних). Таким чином, Agentic RAG мінімізує імовірність пропуску важливої інформації і коригує потенційні помилки або нерелевантні дані до генерації остаточної відповіді. З практичної точки зору, Advanced RAG є простішою у реалізації та менш ресурсозатратною, тоді як Agentic RAG потребує більшого управління (координації агентів, можливої інтеграції додаткових інструментів), але забезпечує найвищу якість результату. В конкретних умовах організації вибір архітектури може залежати від доступних обчислювальних ресурсів і вимог до точності: для швидких відповідей на типові запитання може вистачити Advanced RAG, тоді як для комплексних аудиторських запитів доцільно задіяти Agentic RAG.

ВИСНОВКИ І ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

Проведене дослідження підтвердило гіпотезу про переваги вдосконалених RAG-архітектур над базовою. Naive RAG (лінійна схема “запит → пошук → генерація” без додаткової логіки) продемонструвала найнижчу ефективність: її відповіді були найменш релевантними до корпоративних політик, часто неповними та містили фактичні неточності. Натомість, архітектури Advanced RAG та Modular/Agentic RAG показали значно кращі результати за всіма метриками якості. Запропонований багатозапитний підхід дозволив підвищити повноту і контекстуальну релевантність відповідей, а агентно-модульний підхід додатково покращив точність завдяки динамічному добору інформації. Всі отримані метрики засвідчили, що вдосконалені RAG-системи суттєво перевершують Naive RAG у задачах опрацювання внутрішніх документів з інформаційної безпеки. Таким чином, для розгортання AI-систем в сфері ІБ доцільно обирати архітектури RAG, збагачені механізмами оптимізації пошуку та інтеграції інструментів, оскільки вони забезпечують більш точні, повні й обґрунтовані відповіді на запитання аудиторів і фахівців з інформаційної безпеки.

Перспективи подальших досліджень полягають у розширенні можливостей агентних підходів та інтеграції їх з новими моделями. Зокрема, цікавим напрямом є Active RAG – підхід, який передбачає активне навчання системи вибирати найбільш інформативні фрагменти для ретривалу та генерації [4]. Це може ще більше підвищити ефективність за рахунок адаптивного фокусування на критичних даних. Також планується дослідити застосування запропонованих архітектур на ширшому переліку задач (наприклад, для суміжних сфер контролю відповідності або аналізу ризиків) та оцінити їх стійкість до появи нових документів і змін у політиках безпеки. В цілому, розвиток Retrieval-Augmented Generation у поєднанні з автономними агентами відкриває нові можливості для побудови інтелектуальних інформаційних систем, здатних ефективно використовувати великі обсяги корпоративних знань.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Крет, Т. Б. (2024). Підходи до моделювання загроз під час створення комплексної системи захисту інформації для багаторівневих інтелектуальних систем керування. *Комп'ютерні системи та мережі, 6*(1), 81–88. https://doi.org/10.23939/csn2024.01.081
2. Марценюк, Є. В., Партика, А. І., & Крет, Т. Б. (2025). Дослідження вразливостей штучного інтелекту та побудова комплексної моделі безпеки організації. *Сучасний захист інформації, 1*(61), 206–218. https://doi.org/10.31673/2409-7292.2025.018929
3. Олійник, Б. В., & Чичкарьов, Є. О. (2025). Методи реалізації Retrieval-Augmented Generation у поєднанні з сучасними великими мовними моделями. *Наукові записки Державного університету телекомунікацій. Серія: Інформаційні технології*, (3), 56–65. https://doi.org/10.31673/2412-1034.2025.03.3267
4. Gao, Y., Xiong, Y., Gao, X., et al. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv*. https://doi.org/10.48550/arXiv.2312.10997
5. International Organization for Standardization. (2022). *ISO/IEC 27001:2022 Information technology—Security techniques—Information security management systems—Requirements*. ISO.
6. Jiang, Z., Xu, F. F., Gao, L., et al. (2023). *Active retrieval augmented generation*. arXiv. https://doi.org/10.48550/arXiv.2305.06983

7. Karpukhin, V., Oguz, B., Min, S., et al. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)* (pp. 6769–6781). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.550
8. LangChain. (n.d.). *LangChain AI Python documentation*. https://python.langchain.com/en/latest/
9. Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html
10. Li, Z., Wang, J., Jiang, Z., et al. (2024). *DMQR-RAG: Diverse multi-query rewriting for RAG*. arXiv. https://doi.org/10.48550/arXiv.2411.13154
11. Singh, A., Ehtesham, A., Kumar, S., et al. (2025). *Agentic retrieval-augmented generation: A survey on agentic RAG*. arXiv. https://doi.org/10.48550/arXiv.2501.09136
12. Yao, S., Zhao, J., Yu, D., et al. (2023). ReAct: Synergizing reasoning and acting in language models. In *Proceedings of the 11th International Conference on Learning Representations (ICLR 2023)*. https://doi.org/10.48550/arXiv.2210.03629
13. Zhao, P., Zhang, H., Yu, Q., et al. (2024). *Retrieval-augmented generation for AI-generated content: A survey*. arXiv. https://doi.org/10.48550/arXiv.2402.19473

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 2 (34), 2026
ISSN 2663 – 4023

Andrii Vynar
Master's degree in software engineering
Interregional Academy of Personnel Management, Kyiv, Ukraine
ORCID: 0000-0003-2044-7590
andrii.vynar.om@gmail.com

Oleksii Svedeniuk
PhD Student, Department of Information Security
Lviv Polytechnic National University, Lviv, Ukraine
ORCID: 0009-0001-6125-7382
oleksii.svedeniuk.asp.2025@lpnu.ua

Ostap Medynskyi
PhD Student, Department of Information Security
Lviv Polytechnic National University, Lviv, Ukraine
ORCID: 0009-0003-6582-828X
ostap.medynskyi.asp.2025@lpnu.ua

EVALUATION OF THE EFFECTIVENESS OF RAG ARCHITECTURES FOR AUTOMATION OF INFORMATION SECURITY POLICY ANALYSIS

Abstract. This paper presents an empirical comparison of the effectiveness of three key Retrieval-Augmented Generation (RAG) architectures – the baseline (Naive RAG), the enhanced (Advanced RAG), and the agent-oriented (Agentic RAG). The study aims to analyze their performance in generating accurate and comprehensive answers to queries in the critical field of information security, based on internal corporate policies and procedures. The experiments were conducted using a specialized Python framework built with the LangChain library, the ChromaDB vector database, and a hybrid configuration of local (Ollama: LLaMA 3) and cloud-based (OpenAI ChatGPT-5) large language model platforms. The results demonstrate the significant potential of Advanced and Agentic RAG architectures to improve the accuracy, completeness, and contextual relevance of generated responses. The proposed approach proves effective for automating and simplifying the processing of large volumes of documentation by information security professionals and auditors.
Keywords: Retrieval-Augmented Generation, Agentic RAG, Advanced RAG, LangChain, ChatGPT-5, information security, LLM.

REFERENCES (TRANSLATED AND TRANSLITERATED)

1. Kret, T. B. (2024). Approaches to threat modeling in the creation of a comprehensive information security system for multi-level intelligent control systems. *Computer Systems and Networks*, *6*(1), 81–88. https://doi.org/10.23939/csn2024.01.081
2. Martseniuk, Ye. V., Partyka, A. I., & Kret, T. B. (2025). Study of artificial intelligence vulnerabilities and development of a comprehensive organizational security model. *Modern Information Security*, *1*(61), 206–218. https://doi.org/10.31673/2409-7292.2025.018929
3. Oliinyk, B. V., & Chychkarov, Ye. O. (2025). Methods of implementing retrieval-augmented generation in combination with modern large language models. *Scientific Notes of the State University of Telecommunications. Series: Information Technologies*, (3), 56–65. https://doi.org/10.31673/2412-1034.2025.03.3267
4. Gao, Y., Xiong, Y., Gao, X., et al. (2023). *Retrieval-augmented generation for large language models: A survey*. arXiv. https://doi.org/10.48550/arXiv.2312.10997
5. International Organization for Standardization. (2022). *ISO/IEC 27001:2022 Information security, cybersecurity and privacy protection—Information security management systems—Requirements*. ISO.
6. Jiang, Z., Xu, F. F., Gao, L., et al. (2023). *Active retrieval augmented generation*. arXiv. https://doi.org/10.48550/arXiv.2305.06983
7. Karpukhin, V., Oguz, B., Min, S., et al. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)* (pp. 6769–6781). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.550

8. LangChain. (2025). *LangChain Python documentation*. https://python.langchain.com/en/latest/
9. Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html
10. Li, Z., Wang, J., Jiang, Z., et al. (2024). *DMQR-RAG: Diverse multi-query rewriting for RAG*. arXiv. https://doi.org/10.48550/arXiv.2411.13154
11. Singh, A., Ehtesham, A., Kumar, S., et al. (2025). *Agentic retrieval-augmented generation: A survey on agentic RAG*. arXiv. https://doi.org/10.48550/arXiv.2501.09136
12. Yao, S., Zhao, J., Yu, D., et al. (2023). ReAct: Synergizing reasoning and acting in language models. In *Proceedings of the 11th International Conference on Learning Representations (ICLR 2023)*. https://doi.org/10.48550/arXiv.2210.03629
13. Zhao, P., Zhang, H., Yu, Q., et al. (2024). *Retrieval-augmented generation for AI-generated content: A survey*. arXiv. https://doi.org/10.48550/arXiv.2402.19473

Отримано редакцією журналу / Received: 04.02.26
Прорецензовано / Revised: 16.02.26
Схвалено до друку / Accepted: 24.09.26