

ШТУЧНИЙ ІНТЕЛЕКТ В КОНТЕКСТІ ЦИФРОВОЇ ТРАНСФОРМАЦІЇ: ВИКОРИСТАННЯ RAG СИСТЕМ

Ювженко Д. І.¹; Путренко В. В., д.т.н., проф.²

Американ Юніверситі Київ

Україна, Київ

¹denys.yuvzhenko@auk.edu.ua; ²viktor.putrenko@auk.edu.ua

Анотація. Бурхливий розвиток моделей генеративного штучного інтелекту у поєднанні з іншими підходами до обробки та аналізу даних надає нові можливості для користувачів інформаційних продуктів та сервісів. Технологічний підхід Retrieval-Augmented Generation надає можливості для такого виду інтеграції, що сприяє процесам автоматизації обробки даних та синтезу нових знань. Поєднання двох цих підходів створює нові можливості у взаємодії користувачів з інформаційними системами нового покоління.

Ключові слова: Генеративний штучний інтелект; Retrieval-Augmented Generation; інформаційні системи

Вступна частина. Сучасний процес цифрової трансформації швидко змінює бізнес-процеси та підвищуючи їх ефективність і швидкість. Ключову роль у цьому відіграють технології штучного інтелекту (ШІ), особливо генеративний ШІ, який не лише автоматизує рутинні завдання, але й генерує нові ідеї та контент. Це дає можливість бізнесу швидше адаптуватися до змін, покращувати взаємодію з клієнтами та автоматизувати складні рішення.

Retrieval-Augmented Generation (RAG) [1][2] — це методика, що поєднує пошук інформації з бази даних і генерацію відповідей за допомогою мовних моделей. Спочатку система знаходить релевантні фрагменти інформації, а потім на їх основі генерує змістовні й точні відповіді.

Генеративний ШІ [3], включно з підходом RAG, дозволяє не просто аналізувати дані, а й знаходити релевантну інформацію та створювати на її основі відповіді в реальному часі. RAG системи поєднують пошук і генерацію контенту, що відкриває нові можливості для автоматизації та прийняття обґрунтованих рішень у різних сферах — від освіти до економіки та менеджменту.

Технології ШІ, зокрема RAG, знижують вартість аналітики, зменшують кількість помилок і підвищують продуктивність. Завдяки автоматизації процесів компанії можуть зосередитися на інноваціях і розвитку, замість витратити ресурси на рутинні операції.

Метою роботи є дослідження технології Retrieval-Augmented Generation для автоматизації обробки та генерації інформації. Це передбачає вивчення можливостей поєднання генеративного штучного інтелекту з пошуковими механізмами для покращення взаємодії користувачів із системами та підвищення ефективності обробки даних у різних галузях, таких як бізнес, наука та освіта.

Основна частина. Для об'єктивної оцінки переваг RAG систем та прийняття рішень про необхідність їх впровадження необхідно розглянути наступне:
1. Переваги RAG систем та нові можливості які вони відкривають. Велика варіація задач яка може бути вирішена за допомогою RAG включає але не обмежує:

- Автоматизація технічної підтримки;
- Оптимізація пошуку технічної документації;
- Автоматизоване створення інструкцій та звітів;

- Персоналізовані рекомендаційні системи для електронної комерції;
- Юридичні сервіси та аналіз контрактів;
- Автоматизація підготовки освітніх матеріалів.

2. Порівняння підходів до роботи з іншими рішеннями на базі генеративного штучного інтелекту, таких як технологія асистентів від OpenAI або використання стандартних моделей таких як Gemini від Google AI або ChatGPT [2] та доцільність використання тих чи інших систем в залежності від контексту задачі.

3. RAG на хмарних платформах [1]. У хмарних середовищах, таких як AWS, можна розгорнути RAG системи з використанням таких сервісів, як Amazon SageMaker для навчання та хостингу моделей або Elasticsearch для організації пошукової бази даних. Хмарні рішення мають переваги у масштабованості, швидкості впровадження та інтеграції з іншими бізнес-системами.

4. Локальні RAG системи [4]. Локальні рішення забезпечують контроль над даними і конфіденційністю, що особливо важливо для сфер, які працюють з чутливою інформацією. Локальні системи можуть бути розгорнуті на внутрішніх серверах компанії з використанням проектів з відкритим кодом для пошуку та генерації.

5. RAG на периферійних пристроях [4] (Edge Devices). Рішення на основі RAG можуть бути розгорнуті на Edge-пристроях, таких як Raspberry Pi або Nvidia Jetson Nano, для автономної обробки даних на місці без необхідності підключення до хмари. Це може бути корисно в промислових середовищах, де доступ до інтернету обмежений, або для систем безпеки та моніторингу. Edge RAG використовується для обробки даних локально, що дозволяє значно скоротити затримки та забезпечити приватність даних. Це може бути застосовано для рішень у виробництві, медицині або логістиці.

6. Перевірка правильності відповіді від RAG. Перевірка точності відповідей, згенерованих RAG, є важливим аспектом для бізнесу. Системи можуть використовувати додаткові модулі для валідації інформації, зіставляючи відповіді з відомими джерелами або перевіреними базами даних.

Висновки. Retrieval-Augmented Generation є перспективною технологією, яка дозволяє значно спростити доступ до інформації та автоматизувати процеси в багатьох галузях. Завдяки своїй гнучкості та здатності інтегруватися як у хмарні середовища, так і в локальні системи, RAG стає ключовим інструментом у цифровій трансформації бізнесу. Однією з найбільш значущих переваг RAG є можливість інтеграції з іншими системами штучного інтелекту, такими як системи обробки природної мови (NLP), комп'ютерний зір, та системи машинного навчання для вдосконалення пошукових і генеративних процесів. Це дозволяє створювати більш інтелектуальні системи, які можуть не лише знаходити і генерувати інформацію, але й інтегрувати її з даними з інших джерел і аналізувати контекст. Інтеграція RAG з іншими видами ШІ, наприклад, зі системами рекомендацій або прогнозними моделями, дає можливість значно підвищити якість користувацького досвіду, персоналізуючи надану інформацію або продукти. Наприклад, у сфері медицини це може означати створення систем, які не тільки витягують і генерують відповіді на основі медичних записів, але й прогнозують можливі діагнози або рекомендації на основі індивідуальних даних пацієнтів. Також інтеграція з системами комп'ютерного зору може дозволити автоматизувати аналіз зображень та генерувати текстові інструкції на основі візуальної інформації. Перспективи розвитку RAG передбачають інтеграцію цієї технології у різні галузі, від автоматизації бізнес-процесів до покращення користувацького досвіду в освітніх та медичних системах. Подальші дослідження можуть зосереджуватися на підвищенні точності пошуку та генерації, а також на оптимізації RAG для роботи на пристроях із обмеженими ресурсами, таких як периферійні пристрої (Edge

devices). Це забезпечить більш автономні рішення, які здатні виконувати завдання на місці без необхідності підключення до хмарних сервісів, що підвищить продуктивність і надійність систем у реальному часі.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

- [1] Amazon Web Services. (n.d.). *What is retrieval-augmented generation (RAG)?* AWS. Retrieved October 18, 2024, from <https://aws.amazon.com/what-is/retrieval-augmented-generation>
- [2] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Amodei, D. (2020). *Language Models are Few-Shot Learners*. *Advances in Neural Information Processing Systems*, 33, 1877-1901.
- [3] Lewis, P., Oguz, B., Rinott, R., Riedel, S., & Stenetorp, P. (2020). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*.
- [4] Zhang, C., Anwer, R. M., Abbasi, Q. H., & Imran, M. A. (2022). *Edge Intelligence: Pushing AI to the edge of the network*. *IEEE Access*, 10, 1176-1189.

Yuvzhenko D. I, Putrenko V. V.

Artificial Intelligence in the Context of Digital Transformation: The Use of RAG Systems

Abstract. *The rapid development of generative artificial intelligence models, combined with other approaches to data processing and analysis, offers new opportunities for users of information products and services. The technological approach of Retrieval-Augmented Generation provides capabilities for integration that facilitates the automation of data processing and the synthesis of new knowledge. The combination of these two approaches creates new possibilities for user interaction with next-generation information systems.*

Keywords: *Generative Artificial Intelligence; Retrieval-Augmented Generation; Information Systems.*