https://doi.org/10.31891/2307-5732-2026-361-53
УДК 004.8:004.62:004.9

НИЧ АНДРІЙ
ТОВ "Бембі"
https://orcid.org/0009-0005-7798-5535
e-mail: nychandrii.g@gmail.com
ПРАВОРСЬКА НАТАЛІЯ
Хмельницький національний університет
https://orcid.org/0000-0001-6001-3311
e-mail: margana2000007@gmail.com

RAG (RETRIEVAL-AUGMENTED GENERATION) ЯК НОВА ПАРАДИГМА КОРПОРАТИВНОЇ АВТОМАТИЗАЦІЇ

У статті досліджено Retrieval-Augmented Generation (RAG) як перспективну парадигму побудови корпоративної автоматизації та знаннєво-орієнтованих інформаційних систем. Показано, що стрімке впровадження великих мовних моделей (Large Language Models, LLM) суттєво розширило можливості природномовних інтерфейсів у бізнес-середовищі, однак LLM-centric автоматизація залишається обмеженою низкою фундаментальних проблем, зокрема галюцинаціями, статичністю знань, закладених у параметрах моделей, та недостатньою доменною спеціалізацією для регульованих і бізнес-критичних сценаріїв. Ці обмеження істотно ускладнюють надійне використання автономних LLM у знаннєво-інтенсивних корпоративних процесах, де критичними є точність, трасованість і відповідність вимогам комплаєнсу.

У роботі проаналізовано еволюцію корпоративної автоматизації від правил-орієнтованих систем і класичних підходів машинного навчання до LLM-орієнтованих рішень та обґрунтовано, що RAG є якісним архітектурним зсувом, а не поступовим удосконаленням наявних підходів. Відокремлення зберігання знань від генеративного ядра та інтеграція зовнішніх механізмів пошуку забезпечують контрольований доступ до актуальних корпоративних баз знань, нормативних документів і операційних даних без необхідності перенавчання моделей. У такій архітектурі LLM виконує насамперед роль механізму логічного виведення, тоді як знання залишаються зовнішньо керованими, перевірюваними та постійно оновлюваними.

Окрему увагу приділено сучасним розширенням RAG, зокрема Ontology-Grounded RAG (OG-RAG), Retrieval-to-Augmented Generation (R2AG) та підходам holistic knowledge retrieval, які спрямовані на підвищення семантичної узгодженості між пошуком і генерацією, зменшення фактичних помилок і підвищення надійності систем у складних корпоративних середовищах. Також розглянуто інтеграцію RAG із мультиагентними механізмами оркестрації, що створює передумови для побудови масштабованих, модульних і бізнес-орієнтованих систем штучного інтелекту (AI).

З позиції корпоративних застосувань RAG розглядається як операційна основа знаннєво-орієнтованої автоматизації у сферах аналізу документів, підтримки відповідності регуляторним вимогам, ухвалення управлінських рішень і клієнтської підтримки. Водночас окреслено ключові виклики впровадження RAG, зокрема зростання інфраструктурної складності, затримки та відсутність уніфікованих бізнес-орієнтованих метрик оцінювання. Зроблено висновок, що Retrieval-Augmented Generation є базовою технологією наступного покоління корпоративних систем автоматизації, яка поєднує адаптивність мовних моделей із контрольованим управлінням корпоративними знаннями.

Ключові слова: *RAG, корпоративна автоматизація, LLM, бази даних, сховища даних, управління знаннями, мультиагентні системи.*

NYCH ANDRII
LLC "Bembi"
PRAVORSKA NATALYA
Khmelnitsky national university, Ukraine

RAG (RETRIEVAL-AUGMENTED GENERATION) AS A NEW PARADIGM FOR ENTERPRISE AUTOMATION

This article examines Retrieval-Augmented Generation (RAG) as an emerging paradigm for enterprise-grade corporate automation and knowledge-centric information systems. While the rapid adoption of large language models (LLM) has significantly expanded the capabilities of natural language interfaces in business environments, LLM-centric automation remains constrained by fundamental limitations, including hallucinations, static knowledge embedded in model parameters, and insufficient domain specificity for regulated and high-risk enterprise scenarios. These limitations restrict the reliable use of standalone LLM in knowledge-intensive business processes where accuracy, traceability, and compliance are critical.

The paper analyzes the evolution of corporate automation from rule-based systems and classical machine learning approaches to LLM-oriented solutions and argues that RAG represents a qualitative architectural shift rather than an incremental improvement. By decoupling knowledge storage from the generative model and integrating external retrieval mechanisms, RAG enables controlled access to up-to-date corporate knowledge bases, policy documents, and operational data without requiring model retraining. In this architecture, LLM primarily function as reasoning engines, while authoritative knowledge remains externally managed, auditable, and continuously updated.

Special attention is given to modern RAG extensions, including Ontology-Grounded RAG (OG-RAG), Retrieval-to-Augmented Generation (R2AG), and holistic knowledge retrieval approaches. These methods improve semantic alignment between retrieval and generation, reduce factual inconsistencies, and enhance robustness in complex enterprise environments. The article also considers the integration of RAG with multi-agent orchestration layers, highlighting their role in supporting scalable, modular, and business-oriented AI systems.

From an enterprise perspective, RAG is positioned as an operational foundation for knowledge-centric automation across document analysis, compliance support, decision-making, and customer service. At the same time, key challenges are identified, such as increased infrastructure complexity, latency overhead, and the lack of standardized business-oriented evaluation metrics. Overall, the study positions Retrieval-Augmented Generation as a core enabling technology for the next generation of corporate automation systems that balance adaptability with controlled enterprise knowledge management.

Keywords: *RAG, enterprise automation, LLM, databases, data warehouses, knowledge management, multi-agent systems.*

Стаття надійшла до редакції / Received 02.12.2025
Прийнята до друку / Accepted 11.01.2026
Опубліковано / Published 29.01.2026

Постановка проблеми у загальному вигляді
та її зв'язок із важливими науковими чи практичними завданнями

Епоха цифрової трансформації супроводжується експоненційним зростанням обсягів даних, що зумовлює потребу у постійному розвитку технологій обробки природної мови (Natural Language Processing, NLP). Великі мовні моделі (Large Language Models, LLM), навчені на масштабних текстових корпусах (зокрема GPT-4, LLaMA, Gemini) [1], продемонстрували здатність генерувати зв'язний, контекстуально релевантний і стилістично коректний текст. Це сприяло їх широкому впровадженню в охороні здоров'я, фінансах, наукових дослідженнях і правовій сфері.

У корпоративному середовищі LLM дедалі частіше розглядаються не лише як інструменти генерації тексту, а як універсальні знаннєві інтерфейси. Практика їх використання, однак, виявила низку фундаментальних обмежень, що стримують надійне застосування автономних мовних моделей у знаннєво-інтенсивних завданнях обробки природної мови.

Ключові обмеження використання LLM у корпоративних сценаріях включають:

– проблему галюцинацій (hallucinations) — генерацію фактично некоректної, але правдоподібної інформації. У спеціалізованих доменах, зокрема в юридичних задачах, рівень помилок може сягати 69-88 %, що істотно підриває довіру до систем штучного інтелекту [2; 3];

– статичність знань, зумовлену фіксацією інформації у параметрах моделі на момент навчання. Це ускладнює роботу з актуальними та динамічно змінними корпоративними даними без повторного донавчання (fine-tuning);

– недостатній рівень доменної експертизи, характерний для загальнопризначених моделей і критичний для бізнес-критичних сценаріїв, що вимагають високої точності, відтворюваності та відповідності регуляторним вимогам.

Для подолання зазначених обмежень та підвищення надійності мовних моделей у знаннєво-інтенсивних сценаріях Lewis та співавт. запропонували архітектуру Retrieval-Augmented Generation (RAG) [4]. Підхід RAG поєднує генеративну модель із зовнішнім механізмом пошуку (retrieval), який забезпечує доступ до релевантних фактів із зовнішніх баз знань або корпоративних сховищ даних перед етапом генерації відповіді.

Надання LLM структурованого й перевіреного контексту дозволяє суттєво зменшити рівень галюцинацій, підвищити фактичну точність відповідей і забезпечити актуалізацію знань без повторного навчання моделі. Емпіричні дослідження показують, що RAG-підходи здатні знизити частку помилкових відповідей у середньому на 37,2 % та підвищити показники фактичної узгодженості більш ніж на 24 відсоткові пункти. В окремих діалогових задачах спостерігається зменшення рівня помилок з 68 % до 10 % при одночасному зростанні показника Knowledge F1 score з 17,7 до 26,0 [2].

Таким чином, Retrieval-Augmented Generation трансформує великі мовні моделі з потенційно ненадійних автономних систем у прозорий, обґрунтований і адаптивний механізм корпоративної автоматизації. Метою даної статті є аналіз технічних засад RAG, узагальнення сучасних методологічних удосконалень цього підходу та обґрунтування його ролі як нової парадигми побудови інтелектуальних корпоративних рішень.

Аналіз досліджень та публікацій
Концепція Retrieval-Augmented Generation

Retrieval-Augmented Generation поєднує два базові компоненти — інформаційний пошук і генерацію відповіді мовною моделлю. У межах класичного RAG-конвеєра первинний запит користувача трансформується у векторне представлення (embedding), після чого здійснюється пошук релевантних фрагментів у векторній базі даних із використанням методів інформаційного пошуку та ранжування документів. Отриманий контекст передається до LLM, яка формує відповідь з урахуванням зовнішніх знань [4]. Така архітектура дозволяє відокремити логічне виведення від зберігання знань, що є принципово важливим для корпоративних інформаційних систем.

З бізнес-перспективи ключові переваги Retrieval-Augmented Generation у корпоративних середовищах полягають у такому:

– інтеграція LLM із внутрішніми джерелами знань без повторного навчання. Корпоративні бази знань, нормативні документи та регламентні матеріали можуть оновлюватися незалежно від генеративного ядра, що забезпечує актуальність інформації та знижує витрати на експлуатацію AI-рішень [2];

– підвищення фактичної точності та зменшення галюцинацій. Надання мовній моделі перевіреного контексту з авторитетних джерел знижує ймовірність генерації некоректної або вигаданої інформації у знаннєво-інтенсивних бізнес-процесах;

– прозорість, пояснюваність і трасованість результатів. Відібрані фрагменти можуть бути явно ідентифіковані як джерела відповіді, що створює підґрунтя для аудиту та відповідності регуляторним вимогам;

– гнучкість і зниження вартості супроводу. Актуалізація знань на рівні пошуку усуває необхідність регулярного перенавчання мовних моделей і дозволяє швидко адаптувати систему до змін бізнес-контексту;

Таким чином, пояснюваність і трасованість у RAG є наслідком архітектурного розділення пошуку та генерації, а не окремою постобробкою результатів.

Подальші дослідження показали, що класична архітектура RAG має обмеження, які стають особливо помітними у складних корпоративних сценаріях та зумовлюють розвиток удосконалених retrieval-орієнтованих підходів.

Серед таких напрямів виокремлюється Ontology-Grounded Retrieval-Augmented Generation (OG-RAG), у межах якого пошук доповнюється використанням доменних онтологій і структурованих знань [6]. OG-RAG дозволяє формалізувати семантичні зв'язки між об'єктами корпоративного домену, зменшуючи неоднозначність пошуку та підвищуючи точність отриманого контексту. Для бізнес-сценаріїв це означає можливість інтеграції нормативних моделей, класифікаторів і корпоративних таксономій у RAG-конвеєр.

Іншим напрямом розвитку є Retrieval-to-Augmented Generation (R2AG), який передбачає глибшу інтеграцію результатів пошуку у процес генерації [7]. На відміну від класичного RAG, де контекст використовується переважно як пасивне доповнення запиту, R2AG забезпечує активний вплив retrieval-інформації на різних етапах генерації. Це зменшує семантичний розрив між пошуком і генерацією та підвищує фактичну узгодженість відповідей, що підтверджується зростанням показників Knowledge F1 score і зниженням рівня галюцинацій.

Отже, сучасна концепція Retrieval-Augmented Generation еволюціонує від базового поєднання пошуку й генерації до семантично узгоджених і знаннєво-орієнтованих архітектур. Інтеграція OG-RAG і R2AG формує підґрунтя для створення надійних, масштабованих і бізнес-орієнтованих корпоративних AI-систем, у яких LLM виступає не джерелом знань, а інтелектуальним механізмом їх інтерпретації та використання.

Обмеження класичного підходу Retrieval-Augmented Generation

Незважаючи на суттєві переваги, класичний підхід Retrieval-Augmented Generation має низку обмежень, які стають особливо помітними у складних корпоративних сценаріях:

– залежність якості генерації від ефективності компонента пошуку. Недостатньо релевантні або семантично неповні фрагменти безпосередньо призводять до зниження фактичної точності відповіді навіть за використання потужних мовних моделей [2];

– семантичний розрив між пошуком і генерацією. У класичних RAG-архітектурах отриманий контекст зазвичай інтегрується у запит у статичному вигляді, що не гарантує його повноцінного використання мовною моделлю під час логічного виведення. Це може призводити до втрати критично важливих знань або їх некоректної інтерпретації [7];

– зростання архітектурної складності та затримок. Використання векторних баз даних, багатоступеневих процесів пошуку та зовнішніх джерел знань підвищує вимоги до інфраструктури й ускладнює інтеграцію RAG у наявні корпоративні IT-ландшафти, що потребує балансування між точністю, швидкодією та вартістю експлуатації [8];

– обмежене врахування структурованих знань і бізнес-логіки. Відсутність явної семантичної моделі домену у процесі пошуку ускладнює використання нормативних правил, онтологій і корпоративних таксономій.

Саме ці обмеження зумовили розвиток удосконалених підходів, зокрема OG-RAG, R2AG та holistic knowledge retrieval-орієнтованих архітектур, які розглядаються у наступних розділах статті.

Архітектурні моделі RAG у корпоративних системах

У сучасних корпоративних системах Retrieval-Augmented Generation реалізується у вигляді багаторівневої архітектури, орієнтованої на підтримку складних бізнес-процесів і знаннєво-інтенсивних сценаріїв. Типовий варіант такої архітектури узагальнено на рис. 1.

Рис. 1. Типова архітектура Retrieval-Augmented Generation у корпоративних системах

Архітектура RAG включає кілька функціонально відокремлених рівнів, кожен з яких відповідає за окремий аспект роботи зі знаннями та генерацією відповідей:

– рівень даних охоплює внутрішні документи, корпоративні бази знань, політики, контракти, історичні записи бізнес-процесів, а також структуровані дані з ERP, CRM та інших інформаційних систем. Основним завданням цього рівня є уніфікація різнорідних джерел і забезпечення їх актуальності. Для корпоративного середовища характерна наявність даних у різних форматах (PDF, DOCX, таблиці, бази даних), що зумовлює потребу у кросформатних і мультимодальних підходах до обробки та пошуку інформації [9; 10];

– рівень пошуку відповідає за індексацію та виявлення релевантної інформації, як правило з використанням векторних баз даних і моделей щільного пошуку (dense retrieval). У корпоративних RAG-архітектурах цей рівень часто доповнюється семантичними фільтрами, метаданими та доменними онтологіями, що дозволяє поєднувати статистичні методи пошуку з бізнес-логікою організації. Подальшим розвитком цього підходу є holistic knowledge retrieval, у межах якого пошук здійснюється одночасно з кількох джерел і форматів для формування цілісного контексту відповіді [9];

– генеративний рівень представлений великою мовною моделлю, яка виконує логічне виведення та формує відповіді на основі отриманого контексту. У корпоративних сценаріях LLM розглядається не як автономне джерело знань, а як інтелектуальний інтерпретатор корпоративної інформації. Інтеграція з OG-RAG і R2AG-підходами підвищує семантичну узгодженість між пошуком і генерацією, що є критично важливим для бізнес-рішень із високою ціною помилки;

– рівень оркестрації забезпечує координацію між компонентами пошуку, генерації та зовнішніми сервісами. У сучасних реалізаціях він дедалі частіше реалізується у вигляді мультиагентної архітектури (agent-based architecture), у межах якої окремі програмні агенти відповідають за пошук, перевірку, узгодження та валідацію результатів. Такий підхід дозволяє реалізовувати багатокрокові бізнес-процеси, контролювати якість відповідей і зменшувати ризики помилкових або некоректних рекомендацій.

Запропонований архітектурний поділ відповідає корпоративним вимогам до масштабованості, керованості та інтеграції RAG-систем із наявними IT-ландшафтами. У порівнянні з лінійними RAG-конвеєрами, агентно-орієнтовані та знаннєво-орієнтовані архітектури забезпечують вищий рівень адаптивності та надійності у складних бізнес-сценаріях. Саме поєднання holistic knowledge retrieval, кросформатної обробки даних і рівня оркестрації формує основу для побудови масштабованих корпоративних AI-рішень, здатних підтримувати стратегічні функції організації [11].

Виклад основного матеріалу та результати дослідження

Мета статті

Метою даної рооти є системний аналіз концепції Retrieval-Augmented Generation, еволюції підходів корпоративної автоматизації та сучасних архітектурних модифікацій RAG, орієнтованих на практичне використання в корпоративних інформаційних системах. Також проведене комплексне дослідження Retrieval-Augmented Generation як нової архітектурної парадигми побудови корпоративної автоматизації та знаннєво-орієнтованих інформаційних систем. У роботі ставиться завдання проаналізувати передумови появи RAG у контексті еволюції корпоративних AI-рішень, зумовленої обмеженнями LLM-centric підходів, зокрема проблемами галюцинацій, статичності знань та недостатньої доменної спеціалізації для бізнес критичних і регульованих сценаріїв використання.

Окремою метою дослідження є системний аналіз архітектурних принципів Retrieval-Augmented Generation, включно з відокремленням генеративного ядра від зовнішніх джерел знань, інтеграцією механізмів інформаційного пошуку та використанням корпоративних баз знань, нормативних документів і операційних даних. У статті також ставиться за мету узагальнення сучасних удосконалень RAG-підходів, зокрема Ontology-Grounded RAG, Retrieval-to-Augmented Generation та holistic knowledge retrieval, з позиції підвищення фактичної точності, семантичної узгодженості та надійності корпоративних AI-систем.

Кінцевою метою роботи є обґрунтування Retrieval-Augmented Generation як концептуальної основи наступного покоління корпоративної автоматизації, орієнтованої на підтримку знаннєво-інтенсивниз бізнес-процесів, прийняття управлінських рішень і забезпечення відповідальності регуляторним вимогам. Отримані результати спрямовані на формуваня методологічного підґрунтя для проєктування масштабованих, керованих і бізнес-орієнтованих RAG-систем, здатних інтегруватися і сучасні корпоративні IT-ландшифти.

Еволюція автоматизації корпоративних систем

Еволюція корпоративної автоматизації тісно пов’язана зі змінами бізнес-моделей, зростанням обсягів даних та підвищенням вимог до швидкості й обґрунтованості управлінських рішень. Якщо на ранніх етапах автоматизація була спрямована переважно на оптимізацію рутинних операцій і зниження операційних витрат, то сучасні підходи дедалі більше орієнтуються на підтримку знаннєво-інтенсивних бізнес-процесів, у яких ключову роль відіграє робота з інформацією та контекстом.

Перший етап розвитку корпоративної автоматизації характеризувався домінуванням правил-орієнтованих систем, побудованих на детермінованих бізнес-правилах і формалізованих регламентах. Такі рішення широко застосовувалися в ERP-системах, бухгалтерському обліку та управлінні ресурсами підприємства, забезпечуючи високу передбачуваність і відповідність внутрішнім політикам організацій. З позиції бізнесу їх основною перевагою була стабільність, однак обмежена гнучкість і висока вартість супроводу суттєво стримували масштабування та адаптацію до динамічних змін ринкового середовища.

Наступний етап еволюції пов’язаний із впровадженням методів машинного навчання у корпоративні рішення. Використання статистичних моделей дозволило автоматизувати прогнозування попиту, виявлення аномалій, кредитний скоринг та інші аналітичні завдання, що безпосередньо впливали на бізнес-ефективність.

Водночас у контексті корпоративної автоматизації ці підходи залишалися зосередженими переважно на структурованих даних і не забезпечували повноцінної роботи з текстовими знаннями, які становлять основу внутрішньої документації, контрактів і нормативних політик організацій [5].

Якісний зсув у бізнес-орієнтованій автоматизації відбувся з появою великих мовних моделей. LLM надали корпоративним системам універсальний інтерфейс взаємодії з неструктурованою інформацією, що відкрило нові можливості для автоматизації підтримки клієнтів, аналізу документів, перевірки відповідності регуляторним вимогам і підтримки управлінських рішень. У практичному вимірі такі моделі почали розглядатися як інтелектуальні асистенти та ядро цифрових працівників. Разом із тим використання LLM як автономного ядра автоматизації підвищує ризики помилок і порушень вимог відповідності, що істотно обмежує їх застосування у регульованих і бізнес-критичних корпоративних сценаріях [2; 3].

У відповідь на зазначені обмеження корпоративна автоматизація еволюціонує від LLM-centric (орієнтованих на LLM) підходів до архітектур, у яких знання відокремлюються від генеративного ядра. Такий зсув відображає перехід від моделі, у якій мовна модель виступає неявним сховищем знань, до підходу з керованими зовнішніми джерелами інформації, що концептуально узагальнено на рис. 2.

Рис. 2. Еволюція корпоративної автоматизації: від правил-орієнтованих систем і LLM-centric підходів до RAG-centric архітектур

У цьому контексті Retrieval-Augmented Generation постає як наступний етап розвитку автоматизації, який поєднує можливості мовних моделей із контрольованими корпоративними джерелами знань.

RAG як основа корпоративної автоматизації

У сучасному корпоративному середовищі Retrieval-Augmented Generation виступає не просто як інструмент генерації відповідей, а як інфраструктурна основа знаннєво-орієнтованої автоматизації. RAG забезпечує уніфікований і керований механізм доступу до корпоративних знаннєвих активів, поєднуючи пошук, інтерпретацію та генерацію інформації в межах єдиного процесу. Це дозволяє інтегрувати роботу зі знаннями безпосередньо у бізнес-процеси, де критичними є контекст, обґрунтованість і відтворюваність рішень.

На відміну від LLM-centric рішень, у яких мовна модель виконує роль автономного джерела відповідей, RAG-орієнтована автоматизація ґрунтується на чіткому розмежуванні між генеративним ядром і корпоративними джерелами знань. Такий підхід забезпечує контроль над використовуваною інформацією та створює передумови для практичного впровадження AI-систем у бізнес-критичних і регульованих середовищах.

Основні напрями застосування RAG у корпоративній автоматизації включають:

– інтелектуальну обробку документів. RAG-системи підтримують аналіз контрактів, внутрішніх регламентів, технічної документації та звітів, поєднуючи пошук релевантних фрагментів із їх семантичним узагальненням. На відміну від традиційних систем документообігу, такі рішення дозволяють виконувати логічне виведення, перевірку відповідності нормативним вимогам і виявлення потенційних ризиків у текстових документах;

– підтримку управлінських рішень і бізнес-аналітику. У цьому сценарії RAG поєднує структуровані корпоративні дані з неструктурованими текстовими джерелами, формуючи єдину інформаційну основу для стратегічних і тактичних рішень. Використання доменних онтологій і корпоративних таксономій у межах OG-RAG підвищує узгодженість результатів із внутрішньою бізнес-логікою організації.

– клієнтську та операційну підтримку. RAG-орієнтовані системи забезпечують формування відповідей на основі актуальних корпоративних знань, сервісної документації та технічних інструкцій. Це підвищує якість обслуговування, зменшує навантаження на персонал і забезпечує узгодженість відповідей у масштабі всієї організації;

– реалізацію багатокрокових бізнес-процесів у мультиагентних архітектурах. Поєднання RAG із агентними рівнями оркестрації дозволяє реалізовувати складні робочі процеси з вбудованим контролем якості, перевіркою результатів і узгодженням із корпоративними правилами та політиками.

У практичному вимірі Retrieval-Augmented Generation забезпечує операційну реалізацію знаннєво-орієнтованої корпоративної автоматизації, трансформуючи великі мовні моделі з автономних генераторів тексту у керований інструмент підтримки бізнес-процесів і управлінських рішень.

Порівняння RAG з класичними підходами автоматизації

Порівняльний аналіз Retrieval-Augmented Generation із класичними підходами корпоративної автоматизації дозволяє чітко окреслити відмінності між процедурно орієнтованими системами та сучасними знаннєво-орієнтованими архітектурами. Узагальнену картину таких відмінностей наведено на рис. 3.

Рис. 3. Порівняння класичних, LLM-centric та RAG-centric підходів до корпоративної автоматизації

Правило-орієнтовані системи, що традиційно використовувалися в корпоративному середовищі, забезпечують високу передбачуваність і відповідність формалізованим бізнес-правилам. Водночас вони характеризуються низькою гнучкістю та значними витратами на супровід: будь-які зміни у бізнес-логіці потребують ручного оновлення правил, що обмежує масштабування таких рішень у динамічних корпоративних сценаріях.

Класичні підходи машинного навчання розширили можливості автоматизації за рахунок використання статистичних моделей, здатних адаптуватися до історичних даних. З бізнес-точки зору це дозволило автоматизувати прогнозування, сегментацію клієнтів і виявлення аномалій. Однак у порівнянні з RAG такі підходи залишаються обмеженими у роботі з неструктурованими текстовими знаннями та, як правило, потребують повторного навчання моделей у разі істотних змін у даних або предметній області [5].

Подальший етап еволюції пов'язаний із LLM-орієнтованою автоматизацією, яка надала корпоративним системам універсальний інтерфейс роботи з природною мовою. Проте в корпоративному контексті обмеження LLM-centric підходів безпосередньо трансформуються у фінансові, правові та репутаційні ризики. Відсутність контролю над джерелами знань, схильність до галюцинацій і статичність навченої інформації істотно ускладнюють використання автономних LLM у регульованих галузях і бізнес-критичних процесах [2; 3].

На цьому тлі Retrieval-Augmented Generation займає проміжну, але стратегічно важливу позицію між жорстко формалізованими та повністю автономними AI-системами. На відміну від класичних ML- та LLM-centric рішень, RAG поєднує гнучкість генеративних моделей із контрольованістю корпоративних джерел знань. Актуалізація інформації здійснюється на рівні пошуку, що усуває потребу у повному перенавчанні моделей і знижує експлуатаційні витрати [4].

Крім того, RAG є більш придатним для використання у регульованих корпоративних середовищах, де критичними є трасованість рішень, можливість аудиту та відповідність нормативним вимогам. Саме ці характеристики роблять RAG практично застосовним у сценаріях із високими фінансовими, правовими та репутаційними ризиками, де автономні LLM-підходи залишаються недостатньо надійними.

Таким чином, у порівнянні з правил-орієнтованими системами, класичними підходами машинного навчання та LLM-centric автоматизацією, Retrieval-Augmented Generation забезпечує збалансований компроміс між адаптивністю, знаннєвою обґрунтованістю та контрольованістю. Це робить RAG найбільш придатним підходом для сучасних корпоративних сценаріїв, у яких вимоги до точності, відповідності та довіри є визначальними.

Обмеження та перспективи розвитку

Незважаючи на суттєві переваги, впровадження Retrieval-Augmented Generation у корпоративних системах супроводжується низкою технічних і організаційних викликів, які необхідно враховувати під час проєктування та експлуатації таких рішень.

Ключові обмеження RAG у корпоративних сценаріях включають:

– залежність якості результатів від ефективності компонента пошуку. Недостатньо релевантні, семантично неповні або застарілі фрагменти знань безпосередньо впливають на коректність генерації навіть за використання сучасних великих мовних моделей. Таким чином, retrieval-компонент стає критичною ланкою всієї RAG-системи [2];

– зростання затримок та інфраструктурної складності. Багатоступеневі процеси індексації та пошуку, використання векторних баз даних і зовнішніх сховищ знань підвищують обчислювальні витрати та ускладнюють інтеграцію RAG у наявні корпоративні IT-ландшафти. У бізнес-середовищах із жорсткими вимогами до часу відповіді це потребує компромісу між точністю, швидкодією та вартістю експлуатації [8];

– відсутність уніфікованих бізнес-орієнтованих метрик оцінювання. Традиційні метрики якості тексту (BLEU, ROUGE, Knowledge F1 score) не завжди коректно відображають практичну бізнес-цінність результатів, що ускладнює порівняння альтернативних рішень і прийняття управлінських рішень щодо їх масштабування.

Подолання зазначених обмежень формує основні напрями подальших досліджень і розвитку RAG-підходів.

Перспективні напрями розвитку Retrieval-Augmented Generation включають:

– стандартизацію RAG-архітектур і формування еталонних моделей для корпоративних застосувань, що дозволить спростити інтеграцію RAG у складні організаційні середовища та знизити поріг входу для підприємств.

– розвиток семантично узгоджених retrieval-підходів, зокрема OG-RAG, R2AG та holistic knowledge retrieval, спрямованих на зменшення семантичного розриву між пошуком і генерацією та підвищення фактичної узгодженості відповідей.

– інтеграцію з мультиагентними архітектурами та рівнями оркестрації, що створює передумови для побудови адаптивних і самоконтрольованих AI-систем, здатних підтримувати складні багатокрокові бізнес-процеси з мінімальним втручанням людини.

Таким чином, подальший розвиток Retrieval-Augmented Generation пов'язаний із переходом від експериментальних прототипів до зрілих корпоративних платформ, у яких надійність, прозорість і бізнес-орієнтовані метрики якості стають ключовими критеріями успіху.

Висновки з даного дослідження
і перспективи подальших розвідок у даному напрямі

У результаті проведеного дослідження обґрунтовано, що Retrieval-Augmented Generation (RAG) слід розглядати не як чергове технічне вдосконалення великих мовних моделей, а як якісно новий архітектурний підхід до побудови інтелектуальних корпоративних систем. Показано, що винесення знань за межі параметрів LLM та їх інтеграція через керовані механізми пошуку дозволяють подолати ключові обмеження LLM-орієнтованої автоматизації, зокрема проблему галюцинацій, статичність знань і нестачу доменної експертизи у бізнес-критичних сценаріях.

У статті системно проаналізовано еволюцію корпоративної автоматизації — від правил-орієнтованих систем і класичних підходів машинного навчання до рішень, побудованих на великих мовних моделях. Доведено, що саме RAG забезпечує збалансований компроміс між адаптивністю генеративних моделей і контрольованістю корпоративних джерел знань. На відміну від автономних LLM-підходів, RAG дозволяє забезпечити трасованість результатів, пояснюваність та відповідність регуляторним вимогам, що є критично важливим для корпоративних середовищ із високими фінансовими та правовими ризиками.

Окрему увагу приділено сучасним архітектурним моделям RAG, зокрема Ontology-Grounded RAG (OG-RAG), Retrieval-to-Augmented Generation (R2AG) та holistic knowledge retrieval-підходам. Показано, що їх поєднання з мультиагентними архітектурами та рівнем оркестрації формує підґрунтя для створення масштабованих, модульних і бізнес-орієнтованих платформ штучного інтелекту. У таких системах LLM виконує роль механізму логічного виведення, тоді як корпоративні знання зберігаються, оновлюються та керуються незалежно, що знижує вартість експлуатації та підвищує гнучкість впровадження.

Водночас встановлено, що широке впровадження RAG у корпоративних системах супроводжується низкою викликів, серед яких інфраструктурна складність, зростання затримок і відсутність уніфікованих бізнес-орієнтованих метрик оцінювання якості. Це зумовлює необхідність подальших досліджень, спрямованих на стандартизацію RAG-архітектур, розробку еталонних моделей для корпоративних застосувань та формування методів оцінювання, орієнтованих на реальну бізнес-цінність.

Таким чином, Retrieval-Augmented Generation постає як основа нової парадигми знаннєво-орієнтованої корпоративної автоматизації, у межах якої інтелектуальні системи переходять від ізольованої генерації тексту до керованого використання корпоративних знань. Подальший розвиток RAG очікувано буде пов'язаний із поглибленням мультиагентної оркестрації, семантично узгоджених підходів до пошуку та трансформацією RAG-рішень у зрілі корпоративні платформи, здатні підтримувати стратегічні бізнес-процеси на рівні організацій.

Література

1. Vaswani A., Shazeer N., Parmar N. et al. Attention Is All You Need // Advances in Neural Information Processing Systems. – 2017. – Vol. 30.

2. Wu S., Xiong Y., Cui Y. та ін. Retrieval Augmented Generation for Natural Language Processing: A Survey // Research Square. – 2025. – DOI: 10.21203/rs.3.rs6959723/v1.

3. Marcus G., Davis E. Rebooting AI: Building Artificial Intelligence We Can Trust. – New York: Pantheon Books, 2019. – 272 p.

4. Lewis P., Perez E., Piktus A. та ін. Retrieval Augmented Generation for Knowledge Intensive NLP Tasks // Advances in Neural Information Processing Systems (NeurIPS). – 2020.
5. Mitchell T. Machine Learning. – New York: McGraw-Hill, 1997. – 414 p.
6. Sharma K., Kumar P., Li Y. OG-RAG: Ontology-Grounded Retrieval-Augmented Generation for Large Language Models // *CoRR*. – 2024. – arXiv:2412.15235.
7. Ye F., Li S., Zhang Y., Chen L. R2AG: Incorporating Retrieval Information into Retrieval Augmented Generation // Findings of ACL: EMNLP. – 2024. – P. 11584–11596.
8. Radeva I., Popchev I., Doukovska L., Dimitrova M. Web Application for Retrieval Augmented Generation: Implementation and Testing // Electronics. – 2024. – Vol. 13, No. 7. – Art. 1361. – DOI: 10.3390/electronics13071361.
9. Tong A., Niu X., Liu Z. et al. Holistic Knowledge Retrieval Augmented Generation over Visually Rich Documents // arXiv:2511.20227. – 2025.
10. Nagy A., Spyridis Y., Argyriou V. Cross Format Retrieval Augmented Generation in XR with LLM for Context Aware Maintenance Assistance // arXiv:2502.15604. – 2025.
11. Nguyen T., Chin P., Tai Y.-W. Multi-Agent Retrieval-Augmented Generation via Collaborative Chain-of-Thought Reasoning / *arXiv preprint* arXiv:2505.20096. – 2025.

References

1. Vaswani A., Shazeer N., Parmar N. et al. Attention Is All You Need // Advances in Neural Information Processing Systems. – 2017. – Vol. 30.
2. Wu S., Xiong Y., Cui Y. та ін. Retrieval Augmented Generation for Natural Language Processing: A Survey // Research Square. – 2025. – DOI: 10.21203/rs.3.rs6959723/v1.
3. Marcus G., Davis E. Rebooting AI: Building Artificial Intelligence We Can Trust. – New York: Pantheon Books, 2019. – 272 p.
4. Lewis P., Perez E., Piktus A. та ін. Retrieval Augmented Generation for Knowledge Intensive NLP Tasks // Advances in Neural Information Processing Systems (NeurIPS). – 2020.
5. Mitchell T. Machine Learning. – New York: McGraw-Hill, 1997. – 414 p.
6. Sharma K., Kumar P., Li Y. OG-RAG: Ontology-Grounded Retrieval-Augmented Generation for Large Language Models // *CoRR*. – 2024. – arXiv:2412.15235.
7. Ye F., Li S., Zhang Y., Chen L. R2AG: Incorporating Retrieval Information into Retrieval Augmented Generation // Findings of ACL: EMNLP. – 2024. – P. 11584–11596.
8. Radeva I., Popchev I., Doukovska L., Dimitrova M. Web Application for Retrieval Augmented Generation: Implementation and Testing // Electronics. – 2024. – Vol. 13, No. 7. – Art. 1361. – DOI: 10.3390/electronics13071361.
9. Tong A., Niu X., Liu Z. et al. Holistic Knowledge Retrieval Augmented Generation over Visually Rich Documents // arXiv:2511.20227. – 2025.
10. Nagy A., Spyridis Y., Argyriou V. Cross Format Retrieval Augmented Generation in XR with LLM for Context Aware Maintenance Assistance // arXiv:2502.15604. – 2025.
11. Nguyen T., Chin P., Tai Y.-W. Multi-Agent Retrieval-Augmented Generation via Collaborative Chain-of-Thought Reasoning / *arXiv preprint* arXiv:2505.20096. – 2025.