DOI: https://doi.org/10.36910/6775-2524-0560-2026-62-27
УДК 62-23
Зайцев Олександр Вікторович, д.т.н.
https://orcid.org/0000-0003-2475-3800
Борисов Олег Володимирович, к.т.н.
https://orcid.org/0000-0002-9460-2605
Румянцев Сергій Олександрович, ад'юнкт
https://orcid.org/0009-0001-0340-0080
Воєнна академія імені Євгенія Березняка, м. Київ, Україна

РОЛЬ І ПЕРСПЕКТИВИ ВИКОРИСТАННЯ АРХІТЕКТУРИ ГЕНЕРАЦІЇ З ДОПОВНЕНИМ ПОШУКОМ RAG В ІНТЕРЕСАХ БЕЗПЕКИ ТА ОБОРОНИ

Зайцев О.В., Борисов О.В., Румянцев С.О. Роль і перспективи використання архітектури генерації з доповненим пошуком RAG в інтересах безпеки та оборони. Стаття присвячена комплексному аналізу архітектури Retrieval-Augmented Generation (RAG) як стратегічно важливого напряму розвитку систем штучного інтелекту для сектору безпеки та оборони. У роботі детально розглянуто проблематику обмеженості знань великих мовних моделей LLM (Large Language Model), зокрема феномен «галюцинацій», застарівання тренувальних даних та ризики роботи з чутливою інформацією. Обгрунтовано, що інтеграція зовнішніх верифікованих джерел знань через механізми RAG дозволяє мінімізувати фактологічні помилки, забезпечити актуальність генерації та надати посилання на першоджерела, що є критичним для військової аналітики. Окрему увагу приділено питанням інформаційної безпеки, цифрового суверенітету та відповідності нормативним вимогам при розгортанні AI-систем. Проведено порівняльний аналіз web-орієнтованих рішень (Web-UI) та локальних платформ (на прикладі RAGFlow). Доведено, що для критично важливих застосувань перевагу мають локальні архітектури, які гарантують повний контроль над життєвим циклом даних та унеможливлюють витоки інформації. У статті також досліджено еволюцію RAG від простого пошуку до складних агентних систем. Описано архітектуру RAGFlow Agent як модульного фреймворку, що дозволяє візуально моделювати робочі процеси. Детально проаналізовано ключові компоненти агентної системи: вузли отримання знань, генерації відповідей, класифікації намірів та умовної логіки. Зроблено висновки щодо перспектив впровадження таких систем для автоматизації процесів підтримки прийняття рішень, аналізу розвідувальних даних та роботи з нормативною документацією.
Ключові слова: штучний інтелект, великі мовні моделі, архітектура RAG, RAGFlow, інформаційна безпека, військові технології, автоматизація, агентні системи, підтримка прийняття рішень

Zaitsev O., Borysov O., Rumyantsev S. The role and prospects of using retrieval-augmented generation (rag) architecture for security and defense. The article provides a comprehensive analysis of the Retrieval-Augmented Generation (RAG) architecture as a strategically important direction for the development of artificial intelligence systems in the security and defense sectors. The paper examines the limitations of Large Language Models (LLM), specifically the phenomenon of "hallucinations," the obsolescence of training data, and the risks associated with handling sensitive information. It is substantiated that integrating external verified knowledge sources via RAG mechanisms minimizes factual errors, ensures the relevance of generated content, and provides citations to primary sources, which is critical for military analytics. Special attention is paid to information security, digital sovereignty, and compliance with regulatory requirements during AI system deployment. A comparative analysis of web-oriented solutions (Web-UI) and local platforms (exemplified by RAGFlow) is conducted. The study demonstrates that local architectures are preferable for mission-critical applications, as they guarantee full control over the data lifecycle and prevent information leakage. The article also explores the evolution of RAG from simple retrieval to complex agentic systems. The architecture of the RAGFlow Agent is described as a modular framework that allows for the visual modeling of workflows. Key components of the agentic system are analyzed in detail, including knowledge retrieval nodes, answer generation nodes, intent classification, and conditional logic. Conclusions are drawn regarding the prospects of implementing such systems for automating decision support processes, intelligence data analysis, and regulatory documentation management.
Keywords: artificial intelligence, Large Language Models, RAG architecture, RAGFlow, information security, military technologies, automation, agentic systems, decision support.

Постановка завдання. В умовах сучасних збройних конфліктів ефективність управління військами та якість прийняття командних рішень значною мірою залежать від оперативності обробки великих масивів інформації, до яких належать розвідувальні дані, бойові донесення, нормативно-правові акти та технічна документація. Стрімкий розвиток технологій штучного інтелекту, зокрема поява потужних великих мовних моделей, відкрив нові перспективи для автоматизації цих процесів, однак безпосереднє використання базових нейромереж у секторі безпеки та оборони пов'язане із суттєвими технологічними та безпековими ризиками [1].

Головною проблемою залишається схильність генеративних моделей до формування правдоподібних, але фактично хибних тверджень, що у фаховій літературі отримало назву «галюцинацій». Така особливість є абсолютно неприпустимою при плануванні операцій чи роботі з керівними документами, де точність інформації має критичне значення. Ситуацію ускладнює статичність знань нейромереж, які обмежені датою завершення їхнього навчання та не мають доступу до оперативних змін обстановки у реальному часі. Крім того, використання публічних

хмарних сервісів для обробки службової інформації створює реальні загрози витоку чутливих даних на сервери третіх сторін, що суперечить вимогам інформаційної безпеки та цифрового суверенітету держави [2].

Вирішенням окресленої проблематики може стати впровадження архітектури генерації, доповненої пошуком (RAG), яка дозволяє об'єднати лінгвістичні можливості мовних моделей із надійністю зовнішніх верифікованих баз знань. Метою цієї статті є дослідження ефективності технології RAG для потреб оборони, проведення порівняльного аналізу веб-орієнтованих та локальних інструментальних платформ, а також обґрунтування доцільності розгортання локальних агентних систем для гарантування безпеки даних [3].

Аналіз досліджень. Фундаментальні дослідження у сфері обробки природної мови отримали значний поштовх завдяки появі архітектури Transformer, яка дозволила створювати великі мовні моделі з високою здатністю до узагальнення тексту. Проте наукова спільнота швидко ідентифікувала критичні вразливості таких систем при їх застосуванні у завданнях, що вимагають верифікації фактів. Численні праці, присвячені проблематиці надійності штучного інтелекту, переконливо доводять, що навіть найсучасніші моделі схильні генерувати помилковий контент у випадках, коли їм бракує точної інформації у внутрішніх вагах [4].

Проривним рішенням проблеми обмеженості пам'яті нейромереж стала концепція Retrieval-Augmented Generation, яка передбачає розділення пам'яті системи на параметричну та непараметричну. Такий підхід, запропонований дослідниками лабораторії Facebook AI Research, дозволяє оновлювати знання системи без необхідності ресурсоємного перенавчання самої нейромережі, шляхом простого додавання нових документів до зовнішнього індексу. Сучасні дослідження у цьому напрямі фокусуються на удосконаленні алгоритмів семантичного пошуку та використанні векторних баз даних для підвищення релевантності вибірки контексту [5].

Водночас аналіз існуючих інструментальних рішень свідчить про домінування хмарних архітектур, орієнтованих на комерційний сектор. Більшість популярних фреймворків не враховують специфічні вимоги оборонної сфери щодо ізольованості обчислень. Окремий науковий інтерес становить еволюція підходів від лінійного пошуку до створення автономних агентних систем, здатних самостійно планувати послідовність дій. Питання побудови повністю автономних локальних агентних середовищ, які б поєднували гнучкість оркестрації процесів із суворими вимогами до захисту інформації, у відкритій літературі висвітлено недостатньо, що й зумовлює актуальність даного дослідження [6].

Метою статті. Дослідження ефективності технології RAG для потреб оборони, проведення порівняльного аналізу веб-орієнтованих та локальних інструментальних платформ, а також обґрунтування доцільності розгортання локальних агентних систем для гарантування безпеки даних.

Виклад основного матеріалу. Теоретичний базис архітектури генерації з доповненим пошуком ґрунтується на поєднанні параметричної пам'яті, що зберігається у вагах нейронної мережі, та непараметричної пам'яті, представленої зовнішнім векторним індексом документів. Формалізація цього процесу описується імовірнісною моделлю, де генерація кожного токена залежить як від вхідного запиту, так і від латентного документа, обраного з домену знань. Математично це виражається через маргіналізацію ймовірності вихідної послідовності по всіх документах бази, що дозволяє системі оперувати інформацією, яка не була доступна на етапі навчання моделі [4].

В епоху, що характеризується зростанням кількості витоків даних, посиленням нормативних вимог (наприклад, GDPR, HIPAA) та підвищеним суспільним скептицизмом щодо хмарного штучного інтелекту, питання архітектури розгортання таких систем переходить із площини технічних переваг у площину стратегічної необхідності. Хоча популярні веб-інтерфейси та хмарні API забезпечують швидке розгортання, вони критично залежать від передачі даних користувача на сторонні сервери. Для сектору безпеки та оборони, який регулярно оперує таємною документацією та інтелектуальною власністю, це створює неприйнятні ризики, оскільки навіть за умови шифрування трафіку зберігається загроза зміни політики провайдера або несанкціонованого доступу третіх сторін. Натомість локально розгорнуті системи з відкритим кодом, такі як RAGFlow, дозволяють реалізувати принципи нульової довіри (Zero Trust), гарантуючи, що дані залишаються виключно в межах захищеного периметра організації [5].

Крім безпекових аспектів, локальні RAG-фреймворки забезпечують необхідний рівень операційного контролю та дотримання нормативних вимог. Організації отримують можливість

інтегрувати мовні моделі з внутрішніми системами автентифікації (LDAP, SSO), підключатися до приватних баз даних та впроваджувати спеціалізовані стратегії фрагментації тексту для складних технічних інструкцій, що є недосяжним при використанні закритих API типу «чорна скринька». Важливим фактором також є незалежність від мережевих з'єднань: локальні системи, запущені на спеціалізованому обладнанні, забезпечують передбачувану продуктивність та низьку латентність, що критично для польових центрів прийняття рішень, де затримки чи відсутність зв'язку є неприпустимими. Геополітичні виклики лише підсилюють потребу в цифровому суверенітеті, змушуючи мінімізувати залежність від іноземної хмарної інфраструктури.

Технологічною відповіддю на ці виклики є перехід від простих пошукових механізмів до агентних архітектур, реалізованих у платформі RAGFlow. Це модульний фреймворк оркестрації, що дозволяє моделювати інтелектуальні робочі процеси у вигляді спрямованих графів. Функціонування такої системи забезпечується взаємодією спеціалізованих вузлів. Центральним елементом є вузол отримання знань, який здійснює гібридний пошук у локальній базі, поєднуючи семантичний аналіз із ключовими словами. Отримана інформація обробляється вузлом генерації, де локальна LLM формує відповідь із обов'язковим посиланням на джерела, забезпечуючи прозорість та можливість верифікації [6].

Для вирішення складних аналітичних завдань архітектура передбачає використання вузлів класифікації намірів та умовної логіки. Класифікатор автоматично визначає тип запиту користувача (наприклад, пошук конкретного наказу чи створення узагальненого звіту) і маршрутизує його відповідною гілкою алгоритму. Це дозволяє створювати гнучкі сценарії роботи, де система самостійно приймає рішення щодо необхідності звернення до тих чи інших баз даних залежно від рівня доступу оператора. Така прозорість архітектури, де кожен крок — від індексації до синтезу відповіді — може бути задокументовано та перевірено, робить локальні агентні системи безальтернативним вибором для побудови надійної IT-інфраструктури у військовій сфері.

Теоретичні основи RAG (математичний апарат). Формула ймовірності в архітектурі RAG є ключовим математичним виразом, що формалізує систему, яка генерує відповідь на основі запиту користувача та зовнішніх документів. Ця формула виглядає так:

$$p(a|q) = \sum_{d \in D} p(d|q) \cdot p(a|q,d), \qquad (1)$$

де: q - вхідний запит (наприклад, питання користувача); a - генерована відповідь; D - множина всіх документів у зовнішньому домені знань; d - окремий документ із цієї множини; $p(d|q)$ - ймовірність того, що документ d є релевантним до запиту; $p(a|q,d)$ - ймовірність згенерувати відповідь a за умови наявності запиту q і документа d.

Формула (1) виражає повну ймовірність генерації відповіді a шляхом усереднення за всіма можливими документами у домені. Іншими словами, система не просто вибирає один «найкращий» документ, а враховує всі документи, зважені за їхньою релевантністю. Чим більша ймовірність $p(d|q)$ тим сильніше документ d впливає на кінцеву відповідь.

На практиці перебирати всю множину D (яка може містити мільйони документів) є обчислювально недоцільним. Тому замість повної суми використовують наближення:

$$p(a|q) \approx \sum_{d \in Top-k(q)} p(d|q) \cdot p(a|q,d), \qquad (2)$$

де $Top - k(q)$ k найбільш релевантних документів (наприклад, k=3 або 5).

Це дозволяє зберегти точність при значному зниженні обчислювальних витрат. Ця формула (2) забезпечує статистично обґрунтований спосіб інтеграції зовнішніх знань. Вона дозволяє моделі: уникати галюцинацій, оскільки відповідь базується на реальних документах; бути прозорою — кожен внесок документа можна простежити через $p(d|q)$; адаптуватися до нових даних без повторного навчання, оскільки достатньо оновити корпус D.

По суті, RAG складається з двох основних компонентів: ретранслятора та генератора. Ретранслятор, який часто реалізується як щільний векторний пошуковий механізм типу FAISS або Elasticsearch, кодує запит користувача у вбудований елемент та шукає в домені документів найбільш релевантні уривки. Генератор, зазвичай представлена великою мовною моделлю, обумовлює свою відповідь як вихідним запитом, так і отриманими документами, синтезуючи відповідь, яка

відображає зовнішні докази. Як складна технічна система, RAG-архітектура може бути представлена як сукупність взаємопов'язаних підсистем, кожна з яких реалізує окрему функціональну роль та водночас формує окрему поверхню ризику з погляду інформаційної безпеки. До таких підсистем належать підсистеми зберігання знань, векторизації та індексації, інформаційного пошуку, генерації, оркестрації та агентної логіки, а також інфраструктурний рівень, що включає обчислювальні ресурси, мережу та засоби моніторингу.

На відміну від хмарних сервісів, локальна інсталяція покладає повну відповідальність за захист даних і стійкість системи на організацію-власника, що вимагає чіткого розмежування довірених і недовірених компонентів, формалізації потоків даних та впровадження принципів secure-by-design і zero-trust на всіх рівнях архітектури. Підсистема зберігання знань у локальній RAG-архітектурі є критичним елементом з точки зору конфіденційності, що зумовлює необхідність поєднання класичних механізмів контролю доступу (RBAC, ABAC) із семантично орієнтованими політиками, які визначають допустимість використання фрагментів знань у конкретному контексті. Аналогічно, підсистема векторизації та індексації має розглядатися як повноцінне сховище конфіденційної інформації, яке потребує шифрування та захисту від побічних витоків.

Ретривер у локальній RAG-системі виконує функцію інтелектуального фільтра між корпусом знань і генератором. З одного боку, він підвищує якість і релевантність відповіді, а з іншого — визначає, які саме фрагменти інформації потрапляють у контекст моделі. У системному вимірі це перетворює ретривер на елемент політики інформаційної безпеки, оскільки некоректні налаштування можуть призвести до витоку надлишкової інформації. Тому для локальних архітектур доцільним є впровадження керованих стратегій пошуку, які враховують рівень секретності та актуальність даних.

Генератор на основі LLM в умовах локального розгортання зазвичай функціонує в ізольованому середовищі, що знижує ризики витоку даних, але підвищує вимоги до контролю поведінки моделі, зокрема фільтрації відповідей та логування процесу генерації.

Окрему роль у локальних RAG-системах відіграє підсистема оркестрації та агентної логіки, яка визначає сценарії використання знань та послідовність звернень до компонентів системи. У безпековому контексті агентна оркестрація виступає механізмом реалізації політик доступу на рівні бізнес-логіки, дозволяючи різним класам користувачів отримувати різні траєкторії обробки запитів. Інфраструктурний рівень локального розгортання, що базується на контейнерних технологіях і мікросервісній архітектурі, забезпечує ізоляцію компонентів та спрощення аудиту, дозволяючи інтегрувати RAG-систему з наявними засобами захисту інформації організації.

Узагальнюючи, слід констатувати, що цінність локальної RAG-архітектури полягає у можливості формалізувати та контролювати всі етапи роботи з даними, що робить її придатною для використання в критично важливих середовищах.

Сучасний етап розвитку технології характеризується переходом від ізольованого використання механізмів пошуку до побудови складних багатокрокових сценаріїв взаємодії, що реалізуються за допомогою агентних підходів.

У цьому контексті RAG трансформується з архітектурного шаблону в основу для формалізованих робочих процесів, які моделюють логіку діяльності організації.

З системної точки зору агент є надбудовою над базовими компонентами, яка інкапсулює ціль та правила поведінки. Інтеграція RAG з агентними підходами дозволяє реалізувати детерміновані сценарії використання знань, що є принципово важливим для систем із підвищеними вимогами до надійності. На рівні архітектури такі системи реалізуються у вигляді графів, де кожен вузол відповідає за окрему функцію: аналіз наміру, вибір джерел, генерацію та валідацію.

З погляду інформаційної безпеки агентний підхід істотно підвищує контроль над використанням RAG, оскільки кожен етап робочого процесу може бути обмежений окремими політиками доступу. Наприклад, агент може виконувати попередню класифікацію запиту і визначати дозволеність звернення до певних сегментів домену знань.

Важливою перевагою формалізованих агентних процесів є можливість відтворюваності та аудиту, що дозволяє здійснювати ретроспективний аналіз рішень та відстежувати використані джерела.

Порівняння web-орієнтованих і локальних RAG-реалізацій доцільно здійснювати з урахуванням трьох ключових критеріїв: надійності функціонування, керованості архітектури та рівня контролю над даними. Кожен із цих критеріїв безпосередньо впливає на придатність системи

для використання в критично важливих інформаційних середовищах. Web-орієнтовані рішення, як правило, базуються на хмарних платформах і надають доступ до потужних моделей без значних початкових витрат, проте їхня надійність є похідною від зовнішнього провайдера. Натомість локальні реалізації забезпечують передбачувану поведінку системи в межах контрольованого середовища, що є критичним для оборонної сфери [5].

Детальний порівняльний аналіз характеристик наведено в таблиці 1. Порівняння веб-орієнтованих і локальних реалізацій доцільно здійснювати з урахуванням критеріїв надійності, керованості та контролю над даними, кожен з яких безпосередньо впливає на придатність системи для використання в оборонній сфері.

Таблиця 1.

Види характеристик	**Web-UI (Хмарні рішення)**	**RAGFlow (Локальна платформа)**
Надійність функціонування	Web-орієнтовані RAG-рішення, як правило, базуються на хмарних LLM-платформах і надають користувачеві доступ до потужних моделей та інфраструктури без значних початкових витрат. З позицій надійності такі системи забезпечують високий рівень доступності за рахунок масштабованих дата-центрів і резервування ресурсів. Водночас їхня надійність є похідною від зовнішнього провайдера: перебої в мережевому з'єднанні, зміни в політиці сервісу або обмеження доступу можуть безпосередньо впливати на працездатність прикладної системи.	Локальні RAG-реалізації, навпаки, забезпечують передбачувану поведінку системи в межах контрольованого середовища. Надійність у цьому випадку визначається якістю внутрішньої інфраструктури, але не залежить від зовнішніх мережевих факторів. Це є суттєвою перевагою для сценаріїв, де необхідна безперервна робота навіть за умов обмеженого або нестабільного зв'язку.
Керованість архітектури	З точки зору керованості web-орієнтовані RAG-системи зазвичай пропонують обмежений набір налаштувань і працюють за моделлю «чорної скриньки». Користувач має мінімальний вплив на внутрішню логіку RAG, стратегії фрагментації документів і поведінку генератора. Це ускладнює адаптацію системи до специфічних предметних областей і знижує прозорість прийняття рішень.	Локальні RAG-системи, навпаки, дозволяють повністю контролювати всі рівні архітектури: від вибору моделей і параметрів індексації до реалізації агентних робочих процесів. Такий рівень керованості є критичним для побудови спеціалізованих інтелектуальних помічників, де важливо не лише отримати відповідь, а й гарантувати коректність логіки її формування.
Контроль над даними	У web-орієнтованих RAG-реалізаціях дані, як правило, передаються до зовнішньої інфраструктури, що створює ризики витоку інформації, порушення вимог щодо локалізації даних і залежності від політик постачальника. Навіть за умови формальних гарантій конфіденційності організація втрачає повний контроль над життєвим циклом даних.	Локальні RAG-реалізації забезпечують повний суверенітет даних: усі етапи від зберігання документів і побудови ембеддингів до генерації відповідей відбуваються в межах контрольованої інфраструктури. Це дозволяє реалізувати суворі політики доступу, аудит використання інформації та відповідність національним і галузевим нормативним вимогам.

Порівняння WEB-UI(Web User Interface) та RAGFlow доцільно здійснювати розглядаючи ці платформи не лише як інтерфейси взаємодії з великими мовними моделями, а як комплексні інструменти побудови RAG-систем з різним ступенем керованості, масштабованості та відповідності вимогам інформаційної безпеки.

Обидва рішення орієнтовані на локальне або напівлокальне розгортання, проте відрізняються філософією проєктування та цільовими сценаріями використання. Нижче наведено порівняльну таблицю 2 web-ui та RAGFlow.

Таблиця 2.

Критерій порівняння	**Web-UI (Хмарні рішення)**	**RAGFlow (Локальна платформа)**
Надійність функціонування	Залежить від стабільності зовнішніх серверів та наявності інтернет-з'єднання. Ризики відмов через зміни політик провайдера.	Висока відмовостійкість за рахунок модульності та ізоляції компонентів. Незалежність від зовнішніх каналів зв'язку.
Контроль над даними	Частковий. Дані передаються третім сторонам, що створює ризики витоку конфіденційної інформації.	Повний суверенітет даних. Усі етапи обробки (від індексації до генерації) відбуваються локально в периметрі організації.
Інформаційна безпека	Базовий рівень захисту, що надається провайдером. Обмежені механізми аудиту дій моделі.	Розширені механізми: контроль доступу (RBAC), детальне журналювання, ізоляція в контейнерах, захист від side-channel атак.
Відповідність нормам	Обмежена, часто не відповідає вимогам щодо обробки таємної інформації або персональних даних (GDPR).	Висока. Система може бути налаштована під конкретні вимоги національних регламентів та стандартів безпеки.
Прозорість і аудит	«Чорна скринька». Обмежена можливість трасування джерел знань та логіки прийняття рішень.	Повна трасованість. Можливість аудиту кожного кроку агента, перевірка використаних джерел та логіки розгалуження.
Інтеграція	Обмежена стандартними API або завантаженням файлів. Складнощі з підключенням внутрішніх БД.	Широкі можливості інтеграції з корпоративною інфраструктурою (SQL, NoSQL, LDAP, SSO, внутрішні ERP-системи).
Типові сценарії	Персональні помічники, чат-боти для навчання, PoC (Proof of Concept), некритичні задачі.	Корпоративні аналітичні системи, штабні комплекси, обробка розвідувальних даних, державні реєстри.

З наведеного аналізу випливає, що Web-UI доцільно розглядати виключно як інструмент швидкого доступу до базових можливостей генеративного штучного інтелекту з низькими вимогами до розгортання. У той же час, RAGFlow виступає повноцінною інженерною платформою, орієнтованою на створення керованих, захищених та масштабованих рішень.

Це робить локальні платформи безальтернативним вибором для побудови архітектури інформаційної підтримки в секторі безпеки та оборони, де вимоги до конфіденційності превалюють над зручністю розгортання [6-9].

Нижче наведено приклади застосування RAG систем на практиці.

Приклади застосування RAG систем наведені на рисунках 1-4.

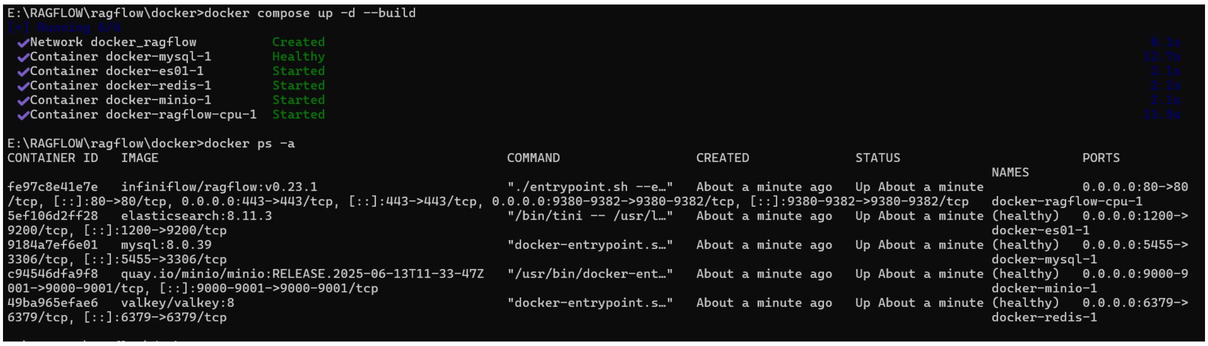

Рис. 1. Запуск інфраструктури на базі контейнерів Docker

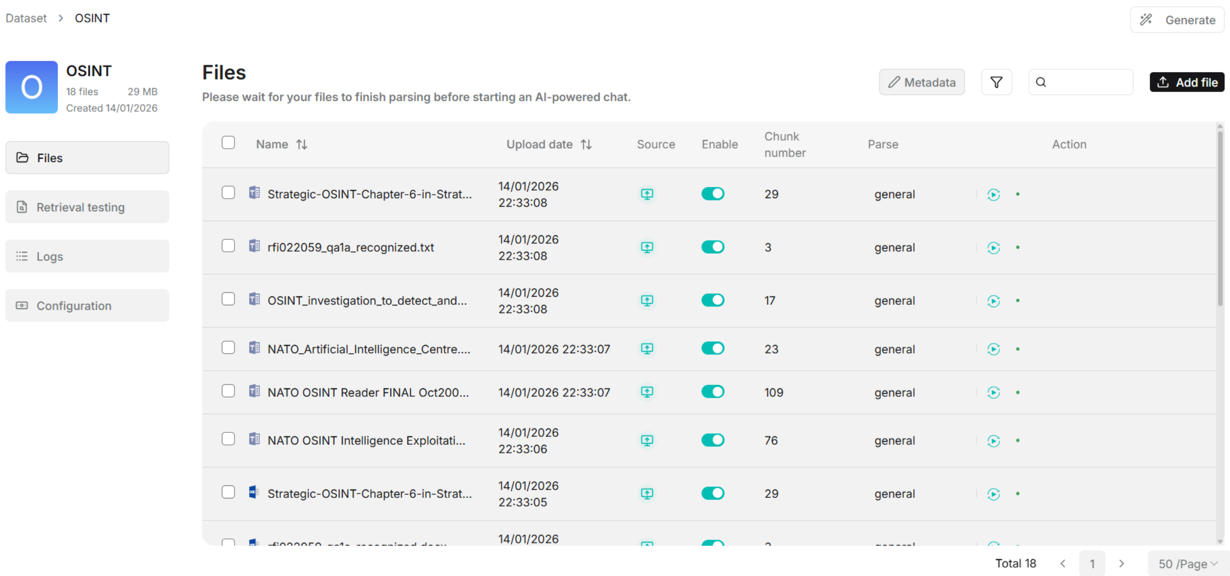

Рис. 2.Приклад Dataset роль OSINT в NATO.

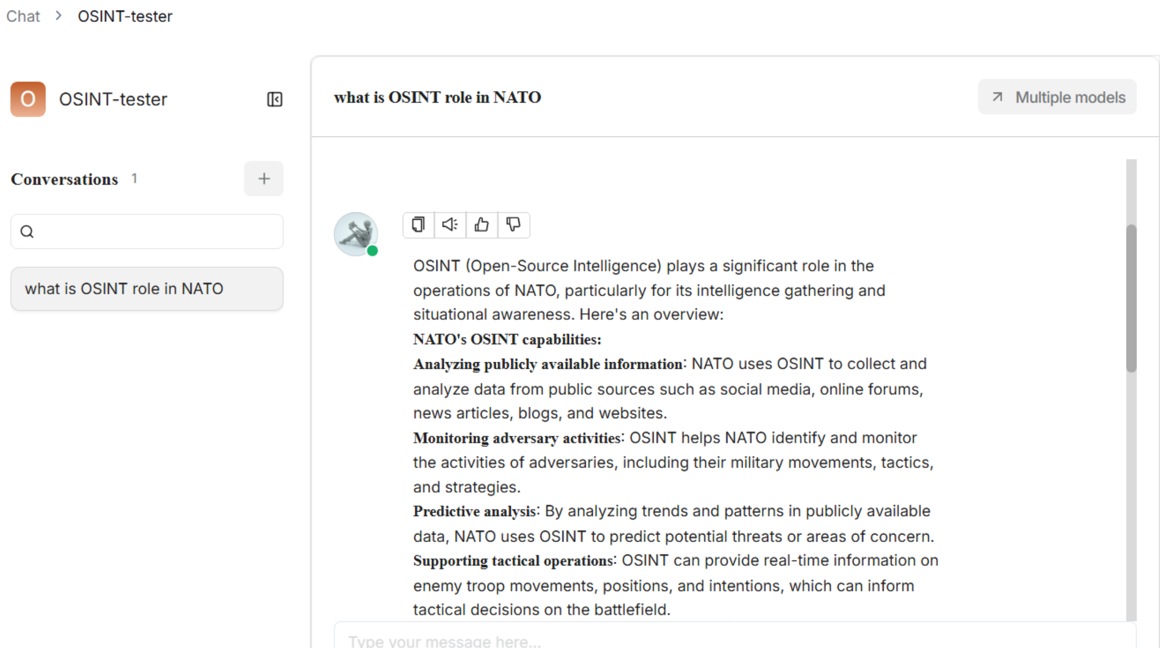

Рис.3. Приклад запиту до LLM з врахуванням і використанням інформації з Dataset

Використання RAG агенту для взаємодії з системою RAGFlow Agent це модульний

фреймворк оркестрації, що дозволяє розробляти, розгортати та керувати інтелектуальними робочими процесами на основі RAG і LLM. Він виконує складні завдання, зокрема підтримку клієнтів, аналіз документів і пошук знань.

Ключові компоненти RAGFlow Agent (рис. 4) вузол отримання знань, що підключається до бази знань і повертає релевантні фрагменти та вузол генерації відповідей і використовує LLM для формування відповідей із посиланням на джерела.

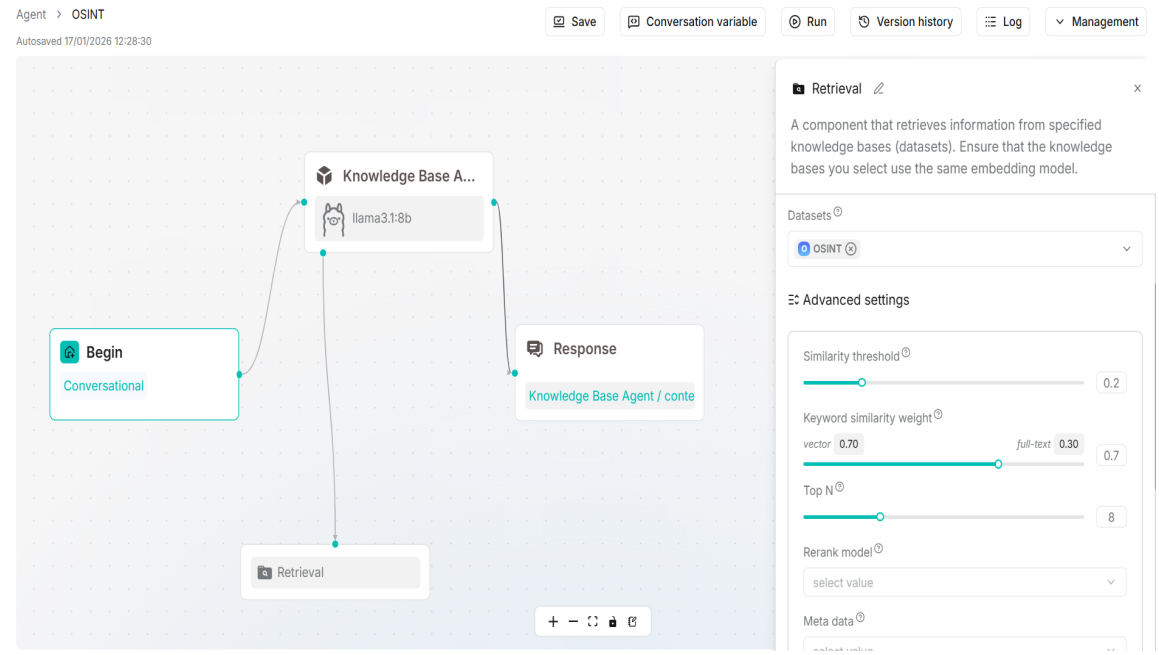

Рис.4. Графова структура агента RAGFlow Agent.

Висновки та перспективи подальшого дослідження. Проведене дослідження дозволяє стверджувати, що архітектура генерації з доповненим пошуком є ефективним технологічним рішенням для подолання фундаментальних обмежень великих мовних моделей, таких як статичність бази знань, схильність до фактологічних галюцинацій та непрозорість механізму прийняття рішень. Інтеграція алгоритмів інформаційного пошуку з генеративним моделюванням на основі розглянутого ймовірнісного апарату забезпечує статистично обґрунтоване використання зовнішніх джерел, що критично підвищує рівень вірогідності сформованих відповідей та робить технологію придатною для використання в контурах управління сектору безпеки та оборони. Критично важливим результатом роботи є обґрунтування необхідності переходу від використання веб-орієнтованих хмарних інтерфейсів до локальних RAG-систем.

Доведено, що умови локального розгортання є обов'язковою вимогою для забезпечення інформаційної безпеки, оскільки вони дозволяють реалізувати повний суверенітет даних, впровадити політики доступу (RBAC/ABAC) та ізолювати обчислювальні процеси від зовнішніх мереж. Порівняльний аналіз показав, що платформи типу RAGFlow забезпечують необхідну модульність та керованість, перетворюючи RAG з простого алгоритму на повноцінну інженерну архітектуру, стійку до кіберзагроз та відмов каналів зв'язку. Еволюція технології від лінійних пошукових алгоритмів до агентних систем, реалізованих через формалізовані робочі процеси, визначає майбутній вектор розвитку галузі.

Використання агентної оркестрації дозволяє моделювати складну логіку використання знань, реалізовувати багатокрокові сценарії аналізу та забезпечувати контроль допустимості інформації на кожному етапі генерації. Такий підхід гарантує відтворюваність результатів та можливість проведення аудиту рішень штучного інтелекту, що є визначальним фактором для легітимізації використання AI-асистентів у штабній роботі. Перспективи подальших наукових розвідок у даному напрямі доцільно зосередити на розробленні формалізованих метрик оцінювання якості роботи RAG-агентів у специфічних предметних областях, створенні методів автоматичної перехресної валідації відповідей кількома моделями, а також на дослідженні можливостей

інтеграції RAG-архітектур із потоковими даними реального часу для підвищення ситуаційної обізнаності.

Список бібліографічного опису

1. Patil G.U., Kwon H.-S., Epureanu B.I., Popa B.-I. Synthetically-trained neural networks for shape classification from measured acoustic scattering // Journal of Sound and Vibration. — 2025. — Vol. 618. — Article 119229. — DOI: https://doi.org/10.1016/j.jsv.2025.119229.
2. Chen Y., Huang Y., Xu C., Zhao X. Multi-fidelity surrogate-based shape optimization using deep neural networks and knowledge-based sampling // Aerospace Science and Technology. — 2024. — Vol. 153. — Article 108932. — DOI: https://doi.org/10.1016/j.ast.2024.108932.
3. Kumar R., Singh R., Sharma M., Patel S. An intelligent fault diagnosis method for roller bearings using ensemble deep learning models // ISA Transactions. — 2024. — Vol. 144. — P. 554–564. — DOI: https://doi.org/10.1016/j.isatra.2024.04.015.
4. Li J., Zhang L., Wang T., Xu K. Damage identification of beam structures using deep autoencoder networks // Structural Control and Health Monitoring. — 2024. — Vol. 31, No. 3. — Art. e3039. — DOI: https://doi.org/10.1002/stc.3039.
5. Park S., Kim Y., Jung H., Lee D. A convolutional neural network-based approach for impact force identification in composite structures // Composite Structures. 2024. Vol. 328. Article 116962. DOI: https://doi.org/10.1016/j.compstruct.2023.116962
6. Wang J., Xu Y., Yang Z., Hu J. A deep learning framework for real-time damage detection of bridges under moving loads // Computer-Aided Civil and Infrastructure Engineering. 2025. Vol. 40, No. 5. P. 482–499. DOI: https://doi.org/10.1111/mice.12997
7. Borisov O., Borisov I. The use of drones to improve the efficiency of communication in combat conditions // Computer-Integrated Technologies: Education, Science, Production. 2023. No. 50. P. 131–135. DOI: https://doi.org/10.36910/6775-2524-0560-2023-50-20
8. Borysov O., Artabaiev Y., Surma A. Cybersecurity of unmanned military aerial vehicles: methods of protection against signal interception and remote control // Computer-Integrated Technologies: Education, Science, Production. 2024. No. 56. P. 117–125. DOI: https://doi.org/10.36910/6775-2524-0560-2024-56-14
9. Zaitsev O., Borysov O. The role and prospects of using robotic complexes in modern combat operations // Computer-Integrated Technologies: Education, Science, Production. 2024. No. 57. P. 174–183. DOI: https://doi.org/10.36910/6775-2524-0560-2024-57-21

References

1. Patil G.U., Kwon H.-S., Epureanu B.I., Popa B.-I. Synthetically-trained neural networks for shape classification from measured acoustic scattering // Journal of Sound and Vibration. — 2025. — Vol. 618. — Article 119229. — DOI: https://doi.org/10.1016/j.jsv.2025.119229.
2. Chen Y., Huang Y., Xu C., Zhao X. Multi-fidelity surrogate-based shape optimization using deep neural networks and knowledge-based sampling // Aerospace Science and Technology. — 2024. — Vol. 153. — Article 108932. — DOI: https://doi.org/10.1016/j.ast.2024.108932.
3. Kumar R., Singh R., Sharma M., Patel S. An intelligent fault diagnosis method for roller bearings using ensemble deep learning models // ISA Transactions. — 2024. — Vol. 144. — P. 554–564. — DOI: https://doi.org/10.1016/j.isatra.2024.04.015.
4. Li J., Zhang L., Wang T., Xu K. Damage identification of beam structures using deep autoencoder networks // Structural Control and Health Monitoring. — 2024. — Vol. 31, No. 3. — Art. e3039. — DOI: https://doi.org/10.1002/stc.3039.
5. Park S., Kim Y., Jung H., Lee D. A convolutional neural network-based approach for impact force identification in composite structures // Composite Structures. 2024. Vol. 328. Article 116962. DOI: https://doi.org/10.1016/j.compstruct.2023.116962
6. Wang J., Xu Y., Yang Z., Hu J. A deep learning framework for real-time damage detection of bridges under moving loads // Computer-Aided Civil and Infrastructure Engineering. 2025. Vol. 40, No. 5. P. 482–499. DOI: https://doi.org/10.1111/mice.12997
7. Borisov O., Borisov I. The use of drones to improve the efficiency of communication in combat conditions // Computer-Integrated Technologies: Education, Science, Production. 2023. No. 50. P. 131–135. DOI: https://doi.org/10.36910/6775-2524-0560-2023-50-20
8. Borysov O., Artabaiev Y., Surma A. Cybersecurity of unmanned military aerial vehicles: methods of protection against signal interception and remote control // Computer-Integrated Technologies: Education, Science, Production. 2024. No. 56. P. 117–125. DOI: https://doi.org/10.36910/6775-2524-0560-2024-56-14
9. Zaitsev O., Borysov O. The role and prospects of using robotic complexes in modern combat operations // Computer-Integrated Technologies: Education, Science, Production. 2024. No. 57. P. 174–183. DOI: https://doi.org/10.36910/6775-2524-0560-2024-57-21

Історія статті:
Отримано: 21.01.2026 Доопрацьовано: 02.03.2026 Прийнято до друку: 23.03.2026 Опубліковано: 29.03.2026