https://doi.org/10.31891/2219-9365-2024-79-19
УДК 681.324

МУСІЄНКО Андрій
Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського"
https://orcid.org/0000-0002-1849-6716
e-mail: mysienkoandrey@gmail.com
ВОРВУЛЬ Данило
Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського"
e-mail: vorvul.danylo@gmail.com
БЕЗСМЕРТНИЙ Владислав
Київський національний університет імені Тараса Шевченка
e-mail: v.y.bessmertniy@gmail.com

# ОЦІНКА ПРОДУКТИВНОСТІ ГОТОВИХ РІШЕНЬ ПОШУКОВО-ДОПОВНЕНОЇ ГЕНЕРАЦІЇ ДЛЯ НАДАННЯ РЕКОМЕНДАЦІЇ З ВИБОРУ НАЙКРАЩОГО РІШЕННЯ

*В даній статті представлено процес розробки застосунка для надання комплексної оцінки існуючих рішень з розширеним пошуком (Retrieval-Augmented Generation) з акцентом на їхню ефективність у фінансовій сфері, зокрема для компаній, що займаються злиттям та поглинанням. Системи RAG, які поєднують механізми пошуку з можливостями генерації, продемонстрували перспективність у наданні точних і контекстно-релевантних відповідей завдяки використанню великих обсягів даних, специфічних для конкретної галузі. Ми створили програмний код, який автоматично оцінює відповіді та дозволяє створити таблицю лідерів для порівняння цих рішень за тринадцятьма ключовими показниками ефективності. Було зібрано датасет з питань та відповідей від дійсних експертів галузі, оцінено здатність кожного рішення обробляти структуровані та неструктуровані фінансові дані. Також порівняно оцінки від експертів та великих мовних моделей, зроблено висновок про ефективність дослідження. Отримані результати висвітлюють сильні і слабкі сторони існуючих систем RAG, надають уявлення про їхню застосовність та потенціал для покращення процесів прийняття рішень у фінансовому секторі. Це дослідження має на меті надання допомоги організаціям у виборі найбільш відповідного рішення RAG для їхніх потреб, а також надати інформацію про майбутні зміни в цій сфері, що швидко розвивається.*

*Ключові слова: Пошуково-доповнена генерація, оцінка ефективності, штучний інтелект, великі мовні моделі, фінансова сфера, інформаційні технології*

MUSIENKO Andrii, VORVUL Danylo, BEZSMERTNIY Vladyslav
National Technical University of Ukraine “Igor Sikorsky Polytechnic Institute”, Taras Shevchenko National University of Kyiv

# EVALUATING THE PERFORMANCE OF OFF-THE-BOX RETRIEVAL AUGMENTED GENERATION SOLUTIONS TO PROVIDE RECOMMENDATIONS ON CHOOSING THE BEST SOLUTION

*This article presents the process of developing an application to provide a comprehensive assessment of existing Retrieval-Augmented Generation (RAG) solutions with a focus on their effectiveness in the financial sector, particularly for M&A companies. RAG systems, which combine search mechanisms with generation capabilities, have shown promise in providing accurate and contextually relevant answers using large amounts of industry-specific data. We created software code that automatically scores the answers and creates a league table to compare these solutions across thirteen key performance indicators. A dataset of questions and answers from actual industry experts was collected, and the ability of each solution to process structured and unstructured financial data was assessed. We also compared the scores from experts and large language models, and concluded on the effectiveness of the study. The findings highlight the strengths and weaknesses of existing RAG systems, provide insight into their applicability and potential to improve decision-making in the financial sector. This study aims to assist organisations in choosing the most appropriate RAG solution for their needs, as well as provide information on future developments in this rapidly evolving field.*

*The study demonstrates the significant potential of advanced search systems in the field of mergers and acquisitions. The results obtained allow organizations to choose the most suitable RAG solution for their needs, increasing the accuracy and relevance of answers to complex queries, which in turn improves decision-making processes in the financial sector.*

*This study makes a significant contribution to the development of knowledge about RAG technologies, especially in the context of their application in the financial sector. The results are an important step in understanding and implementing these technologies, which will help provide more accurate, relevant and useful answers to complex financial queries.*

*Keywords: Search-Augmented Generation, performance evaluation, Artificial Intelligence, Large Language Models, financial sector.*

## ПОСТАНОВКА ПРОБЛЕМИ У ЗАГАЛЬНОМУ ВИГЛЯДІ ТА ЇЇ ЗВ’ЯЗОК ІЗ ВАЖЛИВИМИ НАУКОВИМИ ЧИ ПРАКТИЧНИМИ ЗАВДАННЯМИ

Поява передових технологій штучного інтелекту зробила революцію в багатьох галузях, серед яких фінанси є одним із секторів, що зазнали найбільшого впливу. Однією з передових методологій ШІ, що з'явилася [1-6], є RAG, яка поєднує в собі сильні сторони пошуку інформації та генерації тексту. Системи

RAG використовують специфічні для конкретної галузі набори даних для надання високоточних, контекстно-релевантних відповідей, що робить їх безцінними для завдань, які вимагають тонкого розуміння та точного отримання інформації.

У фінансовій сфері, зокрема у сфері злиття та поглинання, застосунки RAG можуть значно покращити процеси прийняття рішень. Ці системи можуть швидко отримувати та генерувати інформацію з великих масивів фінансових даних, що має вирішальне значення для таких видів діяльності, як аналіз ринку, оцінка ризиків та дотримання нормативних вимог. У контексті злиття та поглинання, де точність і своєчасність мають першорядне значення, системи RAG можуть допомогти в проведенні комплексної перевірки, оціночного аналізу та інтеграційного планування, синтезуючи складну фінансову інформацію, надаючи дієві висновки.

Незважаючи на перспективність рішень RAG, існує нагальна потреба систематично оцінювати їхню ефективність у різних сферах застосування у фінансовій галузі. Існуючі дослідження [7, 8] часто зосереджуються на широких реалізаціях, які не можуть бути застосовані під конкретні юзкейси. Ця стаття має на меті заповнити цю прогалину, розробивши програмну платформу, яка вимірює продуктивність декількох рішень RAG в сфері злиття та поглинання. Створивши стандартизовану таблицю лідерів, ми надаємо порівняльний аналіз цих рішень, виділяючи їхні сильні сторони та сфери, що потребують вдосконалення.

Наше дослідження фокусується на тринадцяти ключових показниках ефективності. Ці показники мають вирішальне значення для оцінки практичної корисності систем RAG у реальних фінансових додатках.

Мета цього дослідження є подвійною. По-перше, ми прагнемо надати чітке, засноване на даних порівняння існуючих рішень RAG, щоб допомогти організаціям вибрати найбільш відповідну систему для їхніх конкретних потреб. По-друге, ми прагнемо виявити обмеження та потенційні можливості для вдосконалення існуючих технологій RAG, тим самим надаючи інформацію для майбутніх досліджень і розробок у цій галузі, що швидко розвивається.

Таким чином, цей документ робить внесок у зростаючий обсяг знань про технології RAG, пропонуючи комплексну систему оцінки, яка охоплює сферу компаній з фокусом на злиття та поглинання. Наші висновки допоможуть зацікавленим сторонам приймати обґрунтовані рішення щодо впровадження та розвитку штучного інтелекту, що в кінцевому підсумку розширить можливості та підвищить ефективність рішень RAG у фінансовому секторі.

## АНАЛІЗ ЛІТЕРАТУРНИХ ДАНИХ ТА ПОСТАНОВКА ПРОБЛЕМИ

Щоденні задачі фінансових експертів полягають у пошуку інформації про компанії та їхній стан, узагальненні цієї інформації та її аналізі. Ефективність цієї роботи безпосередньо впливає на прийняття інвестиційних рішень, розробку фінансових стратегій. Вибір системи для інтеграції в бізнес-процеси компанії є надзвичайно важливим кроком і не може відбуватися без належного етапу тестування. У фінансовій сфері навіть передові великі мовні моделі не можуть досягти точності понад 90% [8]. Тому багато компаній проводять дослідження для своїх специфічних випадків використання. Наразі відсутні дослідження, які б оцінювали ефективність RAG систем у сфері злиттів та поглинань компаній. Проте існує багато досліджень які міряють ефективність великих мовних моделей у відповідях на питання. Зазвичай ці бенчмарки базуються на неспеціалізованих датасетах в конкретній області [9, 10, 11].

Було проаналізовано останнє дослідження, яке максимально близьке до нашого - FinanceBench [8]. Це перший у своєму роді набір тестів для оцінювання ефективності великих мовних моделей з відповіді на фінансові питання. В їх бенчмарку представлено 2 493 питання які передбачають виключно вилучення інформації (28%), 5 897 питань передбачають числові міркування (66%) та 518 (6%) передбачають логічні міркування. На основі цього дослідження можна зробити висновок, що великі мовні моделі набагато гірше справляються з фінансовими питаннями, якщо не мають доступ до додаткової інформації (20% точності проти 79% для GPT-4Turbo). Тому наше дослідження буде базуватись лише на RAG системах. Також FinanceBench дослідження підкреслює лімітації великих мовних моделей при міркуванні з цифрами. Тому при зборі датасету для нашого дослідження було прийнято рішення зосередитись на питаннях з основною задачею пошуку інформації та нескладних логічних міркувань.

Таблиця 1

**Порівняння інструментів для оцінки ефективності систем RAG**

| Фреймворк | Час | Оцінка Метрик | Пошук | Генерація |
|---|---|---|---|---|
| LangChain Benchmark [16] | Листопад 2023 | Точність, Правдивість, Час Виконання, Косинусна Відстань Вбудовування | Точність | ШІ як Оцінювач |
| Databricks Assessment [17] | Грудень 2023 | Коректність, Читабельність, Повнота, Релевантність Контексту, Релевантність Відповіді, Обґрунтованість | - | ШІ як Оцінювач |
| TruEra RAG Group [18] | Жовтень 2023 | Релевантність Контексту, Релевантність Відповіді, Обґрунтованість | ШІ як Оцінювач | ШІ як Оцінювач |

Існує багато інструментів та бенчмарків для оцінки ефективності систем RAG. Зазвичай, оцінювання ефективності поділяється на автоматичне та експертне (проводиться людьми). Автоматичний підхід зазвичай передбачає використання великої мовної моделі як оцінювача, причому цей метод демонструє високу кореляцію з уподобаннями людей (80%) [12, 13]. Основною перевагою автоматичного оцінювання є його здатність до масштабування, оскільки неможливо залучати експертів для оцінки нової версії системи після кожного покращення.

Проте не можна нехтувати оцінкою від експертів, особливо в нашому випадку, де попередніх досліджень ще не проводилося. Це дозволить зрозуміти, наскільки добре система відповідає потребам експертів.

Для автоматичної оцінки було обрано кастомну реалізацію фреймворку RGAR (Retrieval, Generation, Additional Requirement) [13] завдяки його найширшій варіативності використаних метрик. Цей підхід забезпечує всебічний аналіз продуктивності системи, дозволяючи охопити різноманітні аспекти її роботи та забезпечити більш точну та комплексну оцінку рис.1.

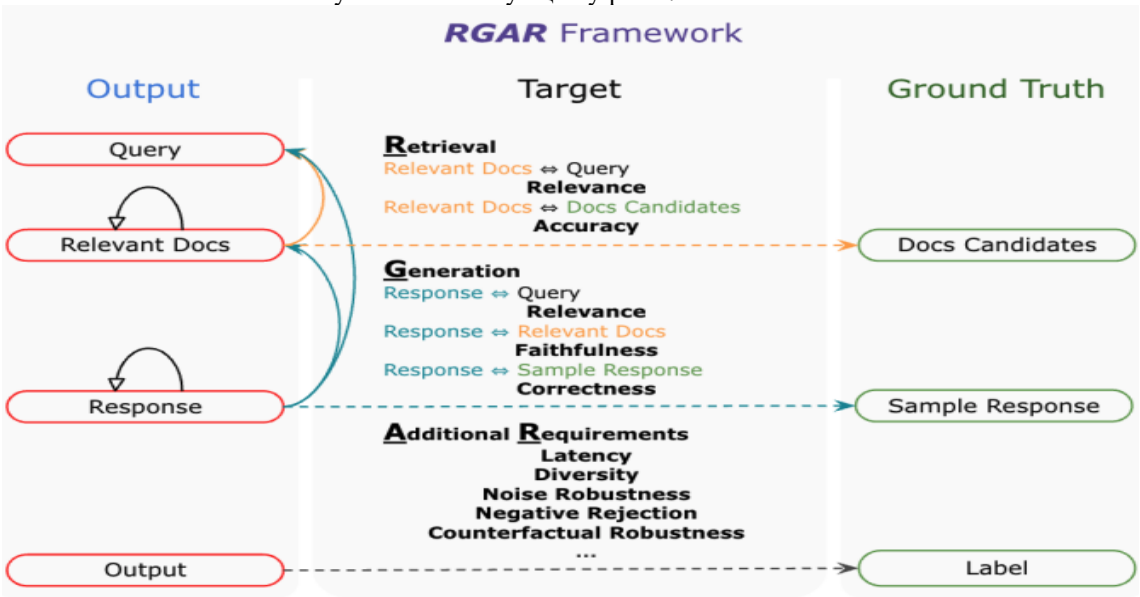


**Рис. 1. Схема реалізації RGAR фреймворку**

Як було встановлено, у відкритому доступі відсутні дослідження, присвячені доменній області компаній, що займаються злиттям та поглинанням. Зазвичай такі компанії при виборі рішень для своїх бізнес-процесів змушені орієнтуватися на продуктивність великих мовних моделей або RAG систем у відповідях на загальні та фінансові питання.

На основі аналізу наявних матеріалів можна окреслити проблему, яка вирішує наше дослідження: відсутність досліджень ефективності RAG систем у сфері діяльності компаній, що займаються злиттям та поглинанням.

**Цілі та завдання дослідження**

Цілями даного дослідження є:

1. Аналіз можливостей систем Retrieval-Augmented Generation (RAG) у наданні точних відповідей на запитання шляхом використання великих мовних моделей для вилучення інформації (включаючи добування певних даних або текстового контенту з файлів) та здійснення нескладних логічних міркувань у домені компаній, що займаються злиттям та поглинанням.
2. Розробка рейтингової таблиці (лідерборду) існуючих рішень RAG. Для досягнення цілей було поставлено наступні завдання:
1. Збір датасету типових запитань та еталонних відповідей для компаній, що займаються злиттям та поглинанням.
2. Налаштування досліджуваних систем, та запуск процесу генерації відповідей на питання.
3. Імплементація фреймворку для оцінювання результатів відповідей системи.
4. Створення лідерборду, де кращі відповіді отримують вищі оцінки.

**Матеріали та методи дослідження**

Для вирішення поставлених задач було створено комплексний додаток для оцінювання відповідей, який інтегрується з наявним програмним забезпеченням на Python. Цей додаток здатний автоматично

оцінювати продуктивність різних систем з розширеним пошуком. Нижче наведено детальну діаграму послідовностей роботи системи:

```
sequenceDiagram
учасник FinancialExpert як Фінансовий експерт
учасник Dataset як Набір даних
      учасник RAG як RAG
      учасник DB як База даних
      учасник EvalApp як Додаток для оцінювання

    Фінансовий експерт->Набір даних: Задати питання
    loop for every система RAG

          Набір даних->RAG: поставити запитання
          RAG->База даних: відповідь у журналі, отримані фрагменти, затримка
          База даних->Додаток для оцінювання: Надання журналів
          Фінансовий експерт->База даних: оцінка згенерованих відповідей
          Додаток для оцінювання->База даних: оцінити отримані фрагменти за метриками P@K, Релевантність, Точність, Джерела

          Додаток для оцінювання->База даних: оцінити згенеровані відповіді за показниками Стійкість до контрфактів, Негативне відхилення, Інтеграція інформації, Контрфактична стійкість, Узгодженість, Релевантність
          Додаток для оцінювання->База даних: Створити звіт на основі зібраних метрик
       кінець
```

**Збір датасету**

Для збору датасету було залучено трьох осіб з досвідом роботи в Private Equity фондах, кожен з яких має відповідний досвід у сфері компаній, що займаються злиттям та поглинанням. Основна діяльність фінансових експертів у цільових компаніях полягає в обробці нотаток щодо угод. Угоди включають декілька стадій; у нашому випадку було розглянуто дві найбільш значущі стадії:

- Approval - остання нотатка перед придбанням компанії, яка використовується для прийняття рішення.
- Closing - стадія, на якій компанія була придбана.

Було зібрано 32 файли з нотатками про угоди купівлі п’яти різних компаній. Для кожної угоди купівлі компанії було зібрано один файл зі стадії Approval, один зі стадії Closing та декілька супутніх матеріалів із додатковою фінансовою інформацією про діяльність компанії. На основі інформації, що розміщена у файлах, експертами було створено 40 питань та надано еталонні відповіді на ці питання.

**Налаштування досліджуваних систем та запуск процесу генерації відповідей на питання**

У цьому розділі ми розглянемо кілька передових RAG-інструментів, зосереджуючи увагу на їхній ефективності у фінансовій сфері. До них належать Azure + AI Search з GPT-3.5/GPT-4, AWS Q, Amazon Bedrock Agents, Plaibox та Nuclia. Кожна з цих систем надає можливості для створення RAG-систем, які слугують базовими рішеннями для подальшої кастомізації та оптимізації під конкретні випадки використання. У кожній з цих систем було створено базу знань та завантажено всі наявні документи. Крім того, написано скрипт, який зчитує датасет із запитаннями, задає їх кожній системі та зберігає відповіді. Розроблений скрипт також зберігає посилання на документи, які були використані для генерації відповідей, та чанки тексту.

**Імплементація фреймворку оцінювання**

Для оцінки якості Retrieval частини було розроблено застосунок який може оцінити відповіді за наступними метриками:

- Міра точності при K (P@K) є одним з показників, що використовується в інформаційному пошуку для оцінки релевантності результатів, які були отримані за запитом або рекомендаційною системою серед перших K позицій. Цей показник є простим, але потужним засобом оцінювання, що дозволяє оцінити точність системи лише на перших K результатах, надаючи інформацію про ефективність системи у виявленні релевантних елементів серед найвищих позицій у її вихідних даних.

Формула для обчислення точності при K (Precision at K, P@K) виглядає наступним чином:

$$P@k(q) = \#\{relevant\ items\ on\ the\ top\ k\ positions\} \div k \quad (1)$$

- Релевантність (Релевантні документи / Запит) оцінює, наскільки знайдені документи відповідають необхідній інформації, вираженій у запиті. Він вимірює точність та специфічність процесу пошуку.
- Точність (Accuracy) є показником, що відображає пропорцію правильних результатів (як істинно позитивних, так і істинно негативних) серед загальної кількості розглянутих випадків. Вона показує, наскільки правильно модель класифікує випадки в різних категоріях, та використовується для оцінки загальної ефективності моделі.

Формула для обчислення точності виглядає наступним чином:

$$Accuracy = \frac{\{TP + TN\}}{\{TP + TN + FP + FN\}} \quad (2)$$

де:

- (TP) (True Positives) - кількість випадків, коли модель правильно визначила позитивний результат.

- (TN) (True Negatives) - кількість випадків, коли модель правильно визначила негативний результат.

- (FP) (False Positives) - кількість випадків, коли модель помилково визначила позитивний результат.

- FN (False Negatives) - кількість випадків, коли модель помилково визначила негативний результат.

- Джерела - наявність посилання на джерело, звідки отримана інформація, щоб можна було швидко перевірити її точність.

Для оцінки Generation частини було обрано наступні метрики спираючись на запропоновані RGAR фреймворку [12]:

- Стійкість до шуму [14, 15]. Цей показник оцінює здатність RAG системи зберігати продуктивність у присутності шуму, нерелевантної або оманливої інформації під час фази пошуку. Модель з високою стійкістю до шуму може ефективно фільтрувати та ігнорувати таку інформацію, зосереджуючись на релевантних даних для генерації точних та зрозумілих відповідей. Стійкість до шуму є критично важливою для забезпечення того, щоб використання зовнішніх механізмів пошуку не знижувало якість результатів моделі.
- Відхилення негативних результатів [14, 15]. Цей показник вимірює здатність RAG системи виявляти та виключати помилково позитивні або неправильно отримані дані, які не відповідають суті запиту. Ця здатність є надзвичайно важливою для запобігання поширенню помилок або дезінформації в згенерованому контенті, забезпечуючи, щоб процес доповнення покращував, а не погіршував якість результатів моделі.
- Інтеграція інформації [14, 15]. Цей показник вимірює, наскільки добре RAG система може синтезувати та включати отриману інформацію з різних джерел у зв'язну та контекстуально відповідну відповідь. Цей показник підкреслює здатність моделі ефективно поєднувати різнорідні дані, що є важливим для створення всебічних та нюансованих відповідей, які відображають глибоке розуміння запиту.
- Стійкість до контрфактів [14, 15]. Стійкість до контрфактів відноситься до здатності LLM обробляти ситуації, що вимагають розуміння та застосування знань, які суперечать загальноприйнятим або очікуваним. Це тестує здатність моделі міркувати та генерувати результати, що відповідно відповідають гіпотетичним або контрфактичним ситуаціям, що свідчить про високий рівень когнітивної гнучкості та креативності.
- Узгодженість [12]. Узгодженість оцінює здатність великої мовної моделі (LLM) створювати зв'язні та логічні відповіді, що є зрозумілими та послідовними. Модель з високим рівнем узгодженості забезпечує, що кожне речення гармонійно вписується в загальний контекст відповіді, не суперечачи одне

одному і утворюючи цілісну картину. Узгодженість є важливою для створення змістовних та зрозумілих відповідей.

- Відповідність [12]. Відповідність вимірює, наскільки добре модель може генерувати інформацію, що безпосередньо стосується заданого запиту. Висока відповідність означає, що модель надає релевантну інформацію, яка точно відповідає на поставлені питання або задовольняє інформаційні потреби користувача. Відповідність критична для забезпечення корисності та точності відповідей.
- Обґрунтованість [12]. Обґрунтованість оцінює, наскільки добре модель базує свої відповіді на достовірних джерелах інформації. Це включає в себе перевірку фактів та надання підтверджувальних доказів для своїх тверджень. Модель з високою обґрунтованістю здатна зменшити ризик поширення дезінформації, забезпечуючи точність та надійність створених відповідей.
- Оцінка людиною. Оцінка людиною передбачає оцінювання відповідей моделі експертами або звичайними користувачами для визначення якості, зрозумілості та точності. Цей метод забезпечує суб'єктивну перевірку продуктивності моделі, враховуючи людське сприйняття і досвід. Оцінка людиною є важливим інструментом для вдосконалення моделі та забезпечення її відповідності очікуванням користувачів. Нашим експертам було запропоновано переглянути отримані відповіді і оцінити їх якість від 1 до 5.
- Затримка. Вимірювалась в секундах в залежності від швидкості конкуруючих систем.

Усі оцінки, за винятком Оцінки людиною, Затримки та оцінок Retrieval, були здійснені за допомогою великих мовних моделей. Для кожної метрики було розроблено відповідний запит, у якому моделі пропонувалося оцінити, наскільки добре оцінка відповідає критерію, за шкалою від 1 до 5.

Приклад промпту:

```
System:
You are an AI assistant. You will be given the definition of an evaluation metric for assessing the quality of an answer in a question-answering task. Your job is to compute an accurate evaluation score using the provided evaluation metric.

User:
Relevance measures how well the answer addresses the main aspects of the question, based on the context. Consider whether all and only the important aspects are contained in the answer when evaluating relevance. Given the context and question, score the relevance of the answer between one to five stars using the following rating scale:
One star: the answer completely lacks relevance
Two stars: the answer mostly lacks relevance
Three stars: the answer is partially relevant
Four stars: the answer is mostly relevant
Five stars: the answer has perfect relevance

This rating value should always be an integer between 1 and 5. So the rating produced should be 1 or 2 or 3 or 4 or 5.

context: Marie Curie was a Polish-born physicist and chemist who pioneered research on radioactivity and was the first woman to win a Nobel Prize.
question: What field did Marie Curie excel in?
answer: Marie Curie was a renowned painter who focused mainly on impressionist styles and techniques.
stars: 1

context: The Beatles were an English rock band formed in Liverpool in 1960, and they are widely regarded as the most influential music band in history.
question: Where were The Beatles formed?
answer: The band The Beatles began their journey in London, England, and they changed the history of music.
stars: 2

context: The recent Mars rover, Perseverance, was launched in 2020 with the main goal of searching for signs of ancient life on Mars. The rover also carries an experiment called MOXIE, which aims to generate oxygen from the Martian atmosphere.
question: What are the main goals of Perseverance Mars rover mission?
answer: The Perseverance Mars rover mission focuses on searching for signs of ancient life on Mars.
stars: 3
```

```
        context: The Mediterranean diet is a commonly recommended dietary plan that
emphasizes fruits, vegetables, whole grains, legumes, lean proteins, and healthy fats.
Studies have shown that it offers numerous health benefits, including a reduced risk
of heart disease and improved cognitive health.
        question: What are the main components of the Mediterranean diet?
        answer: The Mediterranean diet primarily consists of fruits, vegetables, whole
grains, and legumes.
        stars: 4

        context: The Queen's Royal Castle is a well-known tourist attraction in the
United Kingdom. It spans over 500 acres and contains extensive gardens and parks. The
castle was built in the 15th century and has been home to generations of royalty.
        question: What are the main attractions of the Queen's Royal Castle?
        answer: The main attractions of the Queen's Royal Castle are its expansive 500-
acre grounds, extensive gardens, parks, and the historical castle itself, which dates
back to the 15th century and has housed generations of royalty.
        stars: 5

        context: {{context}}
        question: {{question}}
        answer: {{answer}}
        stars:

        Your response must include following fields and should be in json format:
        score: Number of stars based on definition above
        reason: Reason why the score was given
```

**Створення лідерборду**

Для створення лідерборду були використані логи роботи системи. Всі зібрані дані були зведені та обраховані за допомогою пакета pandas, що дозволило зручно організувати інформацію та створити формули для оцінки місця учасників на лідерборді. Цей підхід забезпечив можливість автоматичного оновлення та точного обчислення результатів на основі визначених критеріїв. Кожна метрика має максимальне значення 5 і мінімальне 0. Єдине виключення - метрики точності, але перед підрахунком рядка Totals їх було помножено на 5, щоб всі метрики мали однакову розмірність. Рішення відсортовані по максимальному балу зліва направо див. Таблиця 2. Порівняння кінцевих результатів та порівняння по метриках наведені на рис. 2 та рис. 3 відповідно.

Таблиця 2

**Фінальний лідерборд**

| Competitors | Azure GPT4 | AWS Q | Plaibox (GPT4) | Nuclia (GPT4) | Azure GPT3.5 | AWS Bedrock Agents (Claude 2.1) |
|---|---|---|---|---|---|---|
| **Counterfactual Robustness** | 4,66 | 3,87 | 3,96 | 3,79 | 2,99 | 3,52 |
| **Negative rejection** | 3,74 | 3,92 | 4,29 | 4,2 | 4,65 | 3,83 |
| **Information Integration** | 4,71 | 4,48 | 3,04 | 3,11 | 4,1 | 3,57 |
| **Accuracy %** | 0,91 | 0,83 | 0,83 | 0,48 | 0,73 | 0,66 |
| **P@K** | 0,92 | 0,8 | 0,68 | 0,48 | 0,7 | 0,64 |
| **Coherence** | 4,36 | 3,87 | 2,87 | 2,87 | 3,07 | 2,98 |
| **Relevance** | 4,41 | 3,88 | 3,69 | 3,24 | 3,34 | 2,97 |
| **Groundedness** | 4,5 | 4,35 | 4,05 | 3,06 | 4,05 | 3,74 |
| **Human eval** | 4,43 | 3,39 | 3,91 | 3,48 | 3,48 | 4,17 |
| **Sources** | 4,27 | 2,92 | 2,13 | 4,05 | 3,83 | 3,14 |
| **Latency** | 3,64 | 2,73 | 4,55 | 4,55 | 1,82 | 0,91 |
| **Total** | **47,88** | **41,53** | **40,03** | **38,85** | **38,48** | **35,35** |

**ISSN 2219-9365**

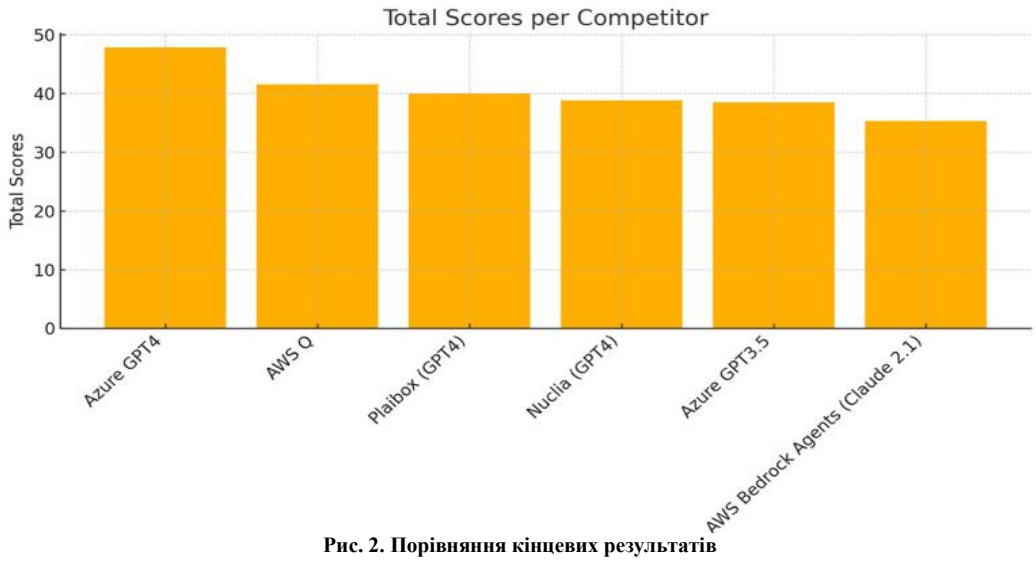


**Рис. 2. Порівняння кінцевих результатів**

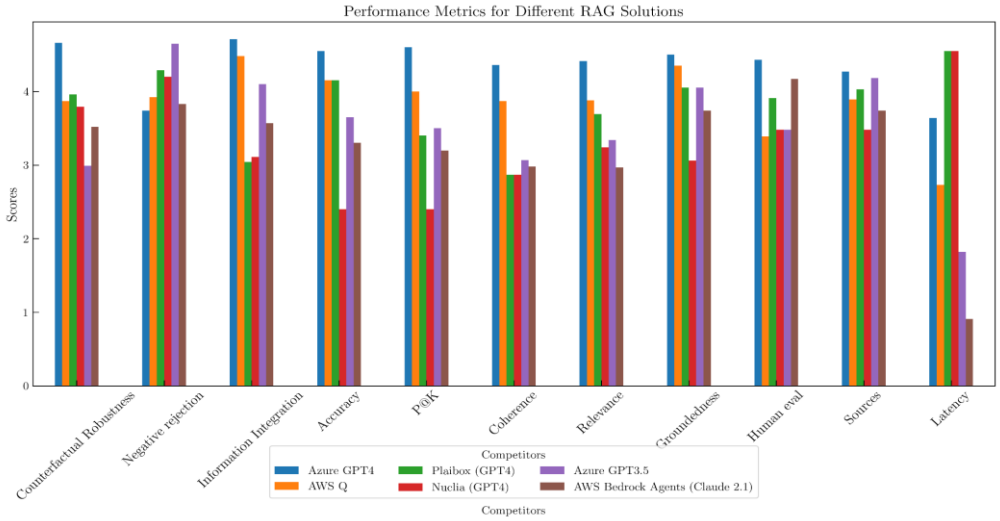


**Рис. 3. Порівняння по метрикам**

Було розраховано кореляцію Пірсона між середнім балом та оцінкою відповідей людиною:

$$r = \frac{\left\{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)\right\}}{\sqrt{\left[n\sum x^2 - \left(\sum x\right)^2\right]\left[n\sum y^2 - \left(\sum y\right)^2\right]}} \tag{3}$$

Ця формула обчислює коефіцієнт кореляції, який коливається від -1 до 1. Значення 1 вказує на ідеальну позитивну кореляцію, -1 вказує на ідеальну негативну кореляцію, а 0 вказує на відсутність кореляції. Кореляція між "Людською оцінкою" та "Середнім балом" становить приблизно 0,382. Це вказує на помірну позитивну кореляцію між цими двома змінними, якщо зібрати більше питань, ми впевнені що кореляція буде тільки збільшуватись.

**Обговорення результатів дослідження**

Отримані результати дослідження свідчать, що серед протестованих рішень найкращі результати показала RAG система Azure OpenAI GPT-4 + Azure AI Search з відривом 15% від другого місця. Всі інші системи працювали на майже однаковому рівні. Ось ключові аспекти її ефективності:

1. Точність: Azure OpenAI GPT-4 + Azure AI Search має найвищий показник точності, що свідчить про її високу здатність надавати правильну інформацію. Це підтверджує її ефективність у фінансовій сфері, де точність даних є критично важливою.

2. Стійкість до контрфактів та інтеграція інформації: Оцінки 4.66 та 4.71 відповідно показують її силу в обробці гіпотетичних сценаріїв та синтезі інформації. Це підкреслює здатність системи справлятися з комплексними завданнями та надавати релевантні відповіді навіть у випадках з наявністю оманливої інформації.

3. Обґрунтованість: Оцінка 4.5 бала відображає ефективність системи у наданні відповідей, підкріплених знайденою інформацією. Це важливо для забезпечення довіри користувачів до отриманих відповідей.

4. Людська оцінка та затримка: Система має хороші показники людської оцінки (4.43) і затримки (3.64), що робить її надійним вибором для додатків, що працюють у режимі реального часу. Це свідчить про те, що результати, надані системою, були високо оцінені експертами, що підвищує її привабливість для кінцевих користувачів.

Також важливо зазначити, що ми отримали помірну кореляцію між людською оцінкою та середнім балом від оцінки великими мовними моделями. Це підтверджує застосовуваність автоматичного підходу до оцінки ефективності RAG систем, дозволяючи ефективно масштабувати процес оцінювання без значних втрат в якості.

Хоча вибірка експертів та питань не є достатньо великою, щоб вважати дослідження фундаментальним для сфери компаній, що займаються злиттям і поглинанням, отримані результати дозволяють застосовувати розроблену методологію для збору більшого датасету і компіляції більш фундаментальних результатів. Це відкриває можливості для подальших досліджень і розширення нашого розуміння ефективності RAG систем у фінансовому секторі. Цей крок був необхідний щоб зумовити необхідність виділення ресурсів на наступну стадію, бо залучення експертів до таких досліджень обходиться дуже дорого.

## ВИСНОВКИ З ДАНОГО ДОСЛІДЖЕННЯ І ПЕРСПЕКТИВИ ПОДАЛЬШИХ РОЗВІДОК У ДАНОМУ НАПРЯМІ

Дослідження демонструє значний потенціал систем з розширеним пошуком у сфері злиття та поглинання компаній. Отримані результати дозволяють організаціям обирати найбільш відповідне рішення RAG для своїх потреб, підвищуючи точність та релевантність відповідей на складні запити, що, в свою чергу, покращує процеси прийняття рішень у фінансовому секторі.

Основні результати:

1. Покращення прийняття рішень: Системи RAG можуть швидко отримувати та генерувати інформацію з великих масивів даних, що є критичним для таких видів діяльності, як аналіз ринку, оцінка ризиків та дотримання нормативних вимог. Це особливо важливо у контексті злиття та поглинання, де точність і своєчасність інформації є ключовими.

2. Стандартизована оцінка ефективності: Створення лідерборду на основі тринадцяти ключових показників ефективності дозволяє чітко порівнювати різні системи RAG. Такий підхід забезпечує комплексну оцінку їх практичної корисності у реальних фінансових додатках, що допомагає виявити сильні та слабкі сторони кожної системи.

3. Інтеграція нових технологій: Дослідження підкреслює важливість інтеграції систем RAG з фінансовими процесами для забезпечення конкурентних переваг на ринку. Використання таких систем дозволяє зменшити час на обробку інформації та підвищити її точність, що є критичним у швидкозмінних ринкових умовах.

4. Можливості для покращення: Виявлені обмеження існуючих систем RAG вказують на напрямки для майбутніх досліджень та розробок. Це стосується покращення обробки специфічних фінансових запитів, точності вилучення інформації, а також інтеграції складних логічних міркувань.

Практичне застосування:

1. Вибір відповідної системи: Результати дослідження надають організаціям цінну інформацію для вибору найбільш відповідного рішення RAG, що підвищує ефективність фінансових аналітиків.

2. Підготовка до майбутніх викликів: Оскільки технології RAG продовжують розвиватися, компанії мають можливість використовувати отримані знання для підготовки до майбутніх викликів та покращення своєї конкурентоспроможності.

Це дослідження робить значний внесок у розвиток знань про технології RAG, особливо в контексті їх застосування у фінансовому секторі. Отримані результати є важливим кроком у розумінні та впровадженні цих технологій, що допоможе забезпечити більш точні, релевантні та корисні відповіді на складні фінансові запити.

**Література**

1. Kwiatkowski, T., et al. (2019). Natural Questions: A Benchmark for Question Answering Research. Transactions of the Association for Computational Linguistics, 7, 453-466.

2. Lewis, P., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv preprint arXiv:2005.11401.

3. Shao, Z., Gong, Y., Shen, Y., Huang, M., Duan, N., & Chen, W. (2023). Enhancing Retrieval-Augmented Large Language Models with Iterative Retrieval-Generation Synergy (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2305.15294.

4. He, X., Tian, Y., Sun, Y., Chawla, N. V., Laurent, T., LeCun, Y., Bresson, X., & Hooi, B. (2024). G-Retriever: Retrieval-Augmented Generation for Textual Graph Understanding and Question Answering (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2402.07630.

5. Mao, Y., Dong, X., Xu, W., Gao, Y., Wei, B., & Zhang, Y. (2024). FIT-RAG: Black-Box RAG with Factual Information and Token Reduction (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2403.14374

6. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W. T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020, May 22). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv.org. https://arxiv.org/abs/2005.11401

7. Gao Y, Xiong Y, Gao X, et al. Retrieval-Augmented Generation for Large Language Models: A survey. arXiv.org. https://arxiv.org/abs/2312.10997. Published December 18, 2023.

8. Islam, P., Kannappan, A., Kiela, D., Qian, R., Scherrer, N., & Vidgen, B. (2023). FinanceBench: A New Benchmark for Financial Question Answering (Версія 1). arXiv. https://doi.org/10.48550/ARXIV.2311.11944

9. Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing Format Boundaries with a Single QA System. In Findings of the Association for Computational Linguistics: EMNLP 2020, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 1896–1907. https://doi.org/10.18653/v1/2020.findings-emnlp.171

10. Matthew Dunn, Levent Sagun, Mike Higgins, V. Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. SearchQA: A New Q&A Dataset Augmented with Context from a Search Engine. arXiv:1704.05179 [cs.CL]

11. Joshua Robinson and David Wingate. 2023. Leveraging Large Language Models for Multiple Choice Question Answering. In The Eleventh International Conference on Learning Representations. https://openreview.net/forum?id= yKbprarjc5B

12. Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena (Version 4). arXiv. https://doi.org/10.48550/ARXIV.2306.05685

13. Yu, H., Gan, A., Zhang, K., Tong, S., Liu, Q., & Liu, Z. (2024). Evaluation of Retrieval-Augmented Generation: A Survey (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2405.07437

14. Chen, M., et al. (2021). Evaluating Large Language Models Trained on Code. arXiv preprint arXiv:2107.03374.

15. Chen, J., Lin, H., Han, X., Sun, L.: Benchmarking large language models in retrieva laugmented generation (Sep 2023). https://doi.org/10.48550/ARXIV.2309. 01431

16. LangChain: Evaluating rag architectures on benchmark tasks (Nov 2023), https://langchain-ai.github.io/langchain-benchmarks/notebooks/ retrieval/langchain_docs_qa.html

17. Leng, Q., Uhlenhuth, K., Polyzotis, A.: Best Practices for LLM Evaluation of RAG Applications (Dec 2023), https://www.databricks.com/blog/ LLM-auto-eval-best-practices-RAG

18. TruLens: TruLens (2023), https://www.trulens.org/trulens_eval/ getting_started/quickstarts/quickstart/

**References**

1. Kwiatkowski, T., et al. (2019). Natural Questions: A Benchmark for Question Answering Research. Transactions of the Association for Computational Linguistics, 7, 453-466.

2. Lewis, P., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv preprint arXiv:2005.11401.

3. Shao, Z., Gong, Y., Shen, Y., Huang, M., Duan, N., & Chen, W. (2023). Enhancing Retrieval-Augmented Large Language Models with Iterative Retrieval-Generation Synergy (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2305.15294.

4. He, X., Tian, Y., Sun, Y., Chawla, N. V., Laurent, T., LeCun, Y., Bresson, X., & Hooi, B. (2024). G-Retriever: Retrieval-Augmented Generation for Textual Graph Understanding and Question Answering (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2402.07630.

5. Mao, Y., Dong, X., Xu, W., Gao, Y., Wei, B., & Zhang, Y. (2024). FIT-RAG: Black-Box RAG with Factual Information and Token Reduction (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2403.14374

6. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W. T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020, May 22). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv.org. https://arxiv.org/abs/2005.11401

7. Gao Y, Xiong Y, Gao X, et al. Retrieval-Augmented Generation for Large Language Models: A survey. arXiv.org. https://arxiv.org/abs/2312.10997. Published December 18, 2023.

8. Islam, P., Kannappan, A., Kiela, D., Qian, R., Scherrer, N., & Vidgen, B. (2023). FinanceBench: A New Benchmark for Financial Question Answering (Версія 1). arXiv. https://doi.org/10.48550/ARXIV.2311.11944

9. Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing Format Boundaries with a Single QA System. In Findings of the Association for Computational Linguistics: EMNLP 2020, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 1896–1907. https://doi.org/10.18653/v1/2020.findings-emnlp.171

10. Matthew Dunn, Levent Sagun, Mike Higgins, V. Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. SearchQA: A New Q&A Dataset Augmented with Context from a Search Engine. arXiv:1704.05179 [cs.CL]

11. Joshua Robinson and David Wingate. 2023. Leveraging Large Language Models for Multiple Choice Question Answering. In The Eleventh International Conference on Learning Representations. https://openreview.net/forum?id= yKbprarjc5B

12. Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena (Version 4). arXiv. https://doi.org/10.48550/ARXIV.2306.05685

13. Yu, H., Gan, A., Zhang, K., Tong, S., Liu, Q., & Liu, Z. (2024). Evaluation of Retrieval-Augmented Generation: A Survey (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2405.07437

14. Chen, M., et al. (2021). Evaluating Large Language Models Trained on Code. arXiv preprint arXiv:2107.03374.

15. Chen, J., Lin, H., Han, X., Sun, L.: Benchmarking large language models in retrieva laugmented generation (Sep 2023). https://doi.org/10.48550/ARXIV.2309. 01431

16. LangChain: Evaluating rag architectures on benchmark tasks (Nov 2023), https://langchain-ai.github.io/langchain-benchmarks/notebooks/ retrieval/langchain_docs_qa.html

17. Leng, Q., Uhlenhuth, K., Polyzotis, A.: Best Practices for LLM Evaluation of RAG Applications (Dec 2023), https://www.databricks.com/blog/ LLM-auto-eval-best-practices-RAG

18. TruLens: TruLens (2023), https://www.trulens.org/trulens_eval/ getting_started/quickstarts/quickstart/