DOI: 10.34185/1991-7848.itmm.2026.01.035

ОЦІНКА ВПЛИВУ ПОПЕРЕДНЬОЇ ФІЛЬТРАЦІЇ НА ЯКІСТЬ ВИБІРКИ В RAG-СИСТЕМАХ З ВЕКТОРНИМ ПОШУКОМ

Клименко І.В.[1] [ORCID], Лебідь Є.А.[2] [ORCID]

[1]УДУНТ, к.е.н., доцент, Україна

[2]УДУНТ, аспірант,Україна

Анотація. *В роботі проаналізовано сучасні підходи до оцінювання RAG-систем, що поєднують векторний пошук і генерацію відповідей великими мовними моделями (LLM). Розглянуто класичні метрики якості вибірки та LLM-орієнтовані метрики якості генерації, у тому числі в контексті фреймворків RAGAS, ARES, VERA та MIRAGE. Проведено машинний експеримент на базі Google Cloud Platform (GCP) Firestore колекції з векторним пошуком по датасету резюме IT-фахівців, де порівнюються стандартний векторний пошук і пошук з попередньою фільтрацією за метаданими. Встановлено, що попередня фільтрація підвищує частку релевантних документів у контексті, зменшує затримку вибірки та дозволяє збільшувати розмір контексту без пропорційного погіршення якості генерації. Результати експерименту підтверджують залежність якості відповідей RAG-систем від чистоти й релевантності контексту.*

Ключові слова: *комп'ютерні системи, інформаційні технології, інтелектуальний аналіз даних, штучний інтелект, RAG, машинний експеримент, генеративні мовні моделі*

Системи Retrieval-Augmented Generation (RAG) поєднують пошук за зовнішніми джерелами з генерацією тексту великими мовними моделями (LLM), що дозволяє підвищувати точність відповіді та опиратися на релевантні для домену дані [1]. Такий підхід використовується в різноманітних комп'ютерних системах – корпоративних асистентах знань, службах підтримки клієнтів, пошуку по технічній документації та інших доменах, де відповідь має ґрунтуватися на заданому корпусі документів. Саме тому комплексна оцінка RAG-систем вимагає застосування окремих метрик для етапів вибірки і формування контексту (retrieval) та генерації відповіді (generation) [2]. У сучасних роботах щодо RAG застосовуються як reference-based (з наявною

множиною еталонних відповідей), так і reference-free (без еталонних відповідей) підходи до оцінювання якості відповіді та контексту.

Мета дослідження – проаналізувати сильні та слабкі сторони різних підходів до оцінки RAG-систем на базі GCP Firestore.

Метрики якості вибірки застосовуються для оцінки того, наскільки добре RAG підбирає контекст для подальшої передачі в LLM. Класичні метрики дають чітку кількісну оцінку якості вибірки, але вимагають наявності еталонної розмітки (множини релевантних документів) для кожного тестового запиту користувача. Така розмітка, як правило, створюється вручну експертами або за допомогою дорогого процесу анотації, що суттєво обмежує можливість підтримки бенчмарків [4]. Фреймворки без застосування еталонних відповідей (наприклад, RAGAS) оцінюють релевантність контексту та коректність відповіді застосовуючи LLM для виділення тверджень (claims) і їх підтвердження контекстом або за допомогою генерації гіпотетичних питань [2]. Перевагою такого підходу є можливість швидко оцінювати інформаційні системи на базі RAG-архітектури без попередньої еталонної розмітки даних.

До класичних метрик якості вибірки відносять:

- Precision@K – визначає частку релевантних документів (або їх фрагментів) серед K результатів пошуку.
- Recall@K – показує, яка частка всіх релевантних документів потрапила до топ-K результатів. Якщо релевантних документів багато, то низьке значення метрики означає втрату важливої інформації для генерації відповіді [4].
- MRR (Mean Reciprocal Rank) – оцінює позицію першого релевантного документу в ранжованому списку. Метрика є актуальною для сценаріїв, де необхідно отримати один релевантний документ на початку списку [3].
- Retrieval Latency (затримка вибірки) – показує час від формулювання запиту користувачем до отримання контексту без урахування часу генерації відповіді.

Варто відмітити, що сучасні фреймворки оцінки RAG (RAGAS, ARES, VERA, MIRAGE та інші) доповнюють класичні метрики якості вибірки додатковими метриками та пропонують різні підходи до їх оцінки.

Метрики якості генерації оцінюють згенеровану LLM відповідь на предмет коректності контексту, релевантності запиту та повноти. Найчастіше використовуються наступні дві метрики:

- Faithfulness – демонструє, наскільки кожне твердження у відповіді узгоджується з наданим LLM контекстом. Значення метрики знаходиться в діапазоні [0, 1], і чим ближче до 1, тим краща узгодженість з контекстом.
- Answer Relevancy – оцінює, наскільки отриманий результат відповідає введеному запиту. Значення метрики коливається від 0 до 1, де вищі показники означають кращу релевантність.

Окремо варто зазначити, що класичні метрики вибірки і генерації застосовуються здебільшого для невеликих контекстів і коротких відповідей. Для RAG-систем із довгими відповідями (long-form RAG), зокрема у завданнях підсумовування великих документів з посиланнями на джерела, потрібно вводити додаткові показники.

Експериментальна частина даного дослідження побудована у GCP Firestore, що підтримує векторний пошук. Тестовий набір даних – резюме кандидатів різноманітних спеціалізацій з IT-сфери Для екстракції і фрагментації тексту, генерації векторних представлень та виокремлення метаданих кожен з документів було автоматично опрацьовано за допомогою моделей сервісу Google Vertex AI. Оброблені дані (текст, векторне представлення, метадані) були збережені в колекції Firestore, що надає можливість застосування стандартного векторного пошуку та пошуку з попередньою фільтрацією за метаданими у межах єдиної колекції. Таблиця 1 демонструє середні значення метрик якості (для різних значень K) та затримки вибірки для обох типів пошуку.

Таблиця 1

Метрики вибірки по типам пошуку

Метрика	Стандартний пошук	Пошук з фільтрацією
Precision@3	0,68	0,89
Recall@3	0,60	0,62
Precision@5	0,56	0,76
Recall@5	0,70	0,69
Precision@10	0,40	0,68
Recall@10	0,85	0,88
MRR	0,94	0,92
Затримка вибірки (мс)	48,91±0,36	19,21±0,19

Як бачимо, у випадку попередньої фільтрації значення метрик суттєво кращі, ніж у випадку стандартного векторного пошуку. Високе значення MRR також свідчить про те, що перший релевантний документ потрапляє на перші позиції видачі. Нижче на рис. 1. зображені метрики якості вибірки (Precision@K та Recall@K) для стандартного пошуку та пошуку з попередньою фільтрацією при різному розмірі контексту (K = 3, 5, 10).

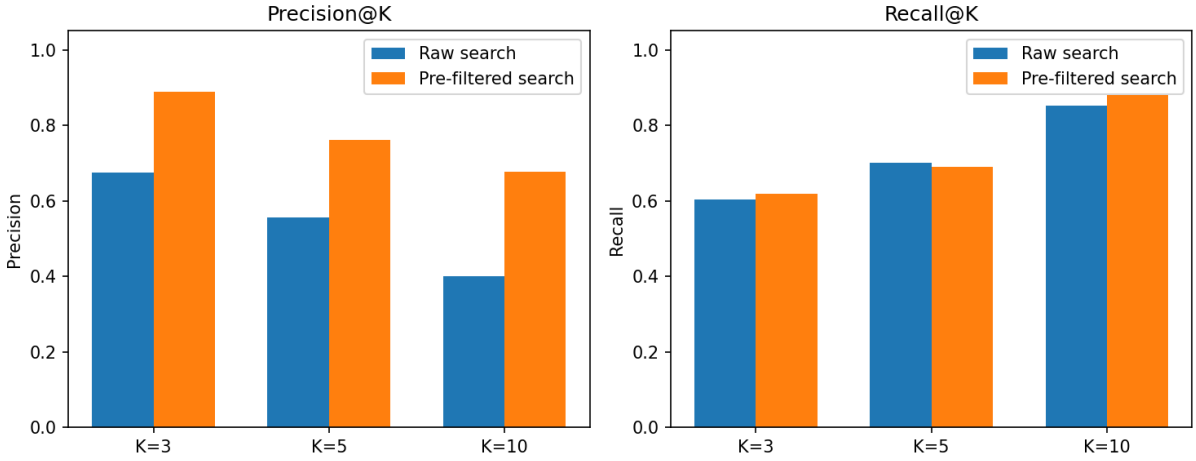

Рисунок 1. Метрики якості вибірки для обох типів пошуку

Пошук з попередньою фільтрацією за метаданими має вищі значення метрики Precision за рахунок зменшення шуму в контексті, а значення показника Recall може змінюватись залежно від того, чи не відсікають фільтри частину релевантних документів. Рис. 2 демонструє середнє значення затримки вибірки (з довірчим інтервалом) для обох типів пошуку.

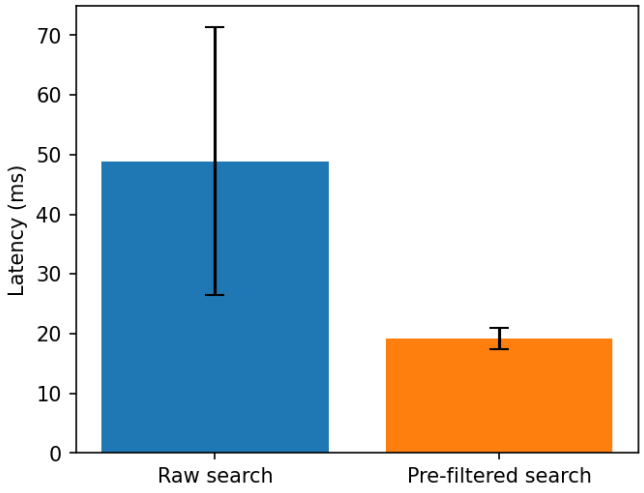

Рисунок 2. Середня затримка вибірки для обох типів пошуку

З графіку видно, що пошук з попередньою фільтрацією значно швидший за стандартний пошук завдяки суттєвому зменшенню об'єму даних для виконання векторного пошуку.

Для оцінки якості генерації проведено компактний експеримент, в якому кожен з тестових запитів проходив через RAG-пайплайн, після чого окрема модель-суддя (LLM-as-judge) оцінила метрики Faithfulness та Answer Relevancy за схемою, близькою до описаної в RAGAS [2]. Експеримент проведено для обох типів векторного пошуку при різному розмірі контексту (K=3 і K=5). В таблиці 2 наведено середні якість генерації для обох типів пошуку.

Таблиця 2

Метрики генерації по типам пошуку

Метрика	Стандартний пошук	Пошук з фільтрацією
Faithfulness (K=3)	1,00	1,00
Answer Relevancy (K=3)	0,94	0,78
Faithfulness (K=5)	0,72	0,94
Answer Relevancy (K=5)	0,65	0,80

У випадку стандартного векторного пошуку середнє значення метрики Faithfulness знижується при збільшенні контексту (K=5) – в контексті з'являється більше даних, тому відповіді можуть містити більше тверджень, частину з яких LLM-суддя не вважає повністю підтвердженою. Для пошуку з попередньою фільтрацією цей показник вищий, ніж у стандартного пошуку за рахунок більш релевантного контексту після фільтрації метаданих. Середні значення показника Answer Relevancy для стандартного пошуку при збільшенні

контексту також знижуються. Зазначимо, що отримане значення Answer Relevancy для пошуку з попередньою фільтрацією вище – це підтверджує гіпотезу про те, що попередня фільтрація зменшує шум у контексті.

Встановлено, що фреймворки на кшталт RAGAS, ARES, VERA, MIRAGE доповнюють класичні метрики якості вибірки та генерації і пропонують різні підходи до їх вимірювання. В той же час, через відсутність єдиного стандарту оцінки RAG [4], експериментальна верифікація окремих рекомендацій є особливо актуальною. На основі проведеного експерименту було з'ясовано наступне:

1. Векторний пошук з попередньою фільтрацією підвищує значення метрики Precision при практично незмінних Recall і MRR порівняно зі стандартним векторним пошуком.
2. Контекст у випадку пошуку з попередньою фільтрацією покращує значення показників Faithfulness та Answer Relevancy, що узгоджується з постулатом про залежність якості генерації від чистоти контексту.
3. При пошуку з попередньою фільтрацією затримка вибірки знижується завдяки зменшенню обсягу даних для виконання операції векторного пошуку на колекції Firestore.
4. Недостатній обсяг даних, переданих судді для валідації, може призводити до помилково низьких оцінок Faithfulness. Дана технічна проблема в проведеному експерименті вирішувалась збільшенням контексту, що передається LLM-судді (з 1000 до 10000 символів).
5. Збільшення контексту покращує Recall, але при стандартному векторному пошуку може погіршувати Faithfulness та Answer Relevancy. Пошук з попередньою фільтрацією дозволяє використовувати більший об'єм контексту без пропорційного погіршення метрик генерації.

ЛІТЕРАТУРА / REFERENCE

1. Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Küttler H., Lewis M., Yih W., Rocktäschel T., Riedel S., Kiela D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems. – 2020. Vol. 33. P. 9459-9474. DOI: 10.48550/arXiv.2005.11401
2. Es S., James J., Espinosa-Anke L., Steven S. RAGAS: Automated Evaluation of Retrieval Augmented Generation. Computer Science. Computation and Language. – 2023. DOI: 10.48550/arXiv.2309.15217
3. Saad-Falcon J., Khattab O., Potts C., Zaharia M. ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). 2024. P. 3464-3483. DOI: 10.48550/arXiv.2311.09476

4. Yu Z., Gan Z., Zhang Y., Tong X., Liu H., Liu Q. Evaluation of Retrieval-Augmented Generation: A Survey. Computer Science. Computation and Language. – 2024. DOI: 10.48550/arXiv.2405.07437

THE ASSESSMENT OF IMPACT OF PRE-FILTERING ON RETRIEVAL QUALITY IN RAG SYSTEMS WITH VECTOR SEARCH

I.V. Klymenko, Y.A. Lebid

Abstract. *The paper analyzes modern approaches to evaluating Retrieval-Augmented Generation (RAG) systems that integrate vector search with answer generation by large language models (LLMs). It examines classical retrieval quality metrics alongside LLM-oriented generation quality metrics, including their application within frameworks such as RAGAS, ARES, VERA, and MIRAGE. A computational experiment was conducted using a Google Cloud Platform (GCP) Firestore collection with vector search over a dataset of IT professionals' CVs, comparing standard vector search against search enhanced by pre-filtering on metadata. The results demonstrate that pre-filtering increases the proportion of relevant documents in the context, reduces retrieval latency, and enables larger context sizes without proportional degradation in generation quality. The experimental findings confirm the dependence of RAG system answer quality on the purity and relevance of the retrieved context.*

Keywords: *computer systems, information technologies, data mining, artificial intelligence, RAG, machine-based benchmarking, generative language models.*