«Системні технології» 3 (164) 2026 «System technologies»
DOI 10.34185/1562-9945-3-164-2026-10
УДК: 004.89; 004.65

І.В. Клименко, Є.А. Лебідь

ОГЛЯД СУЧАСНИХ ФРЕЙМВОРКІВ ТА МЕТРИК ОЦІНКИ RAG-СИСТЕМ

Анотація. Актуальність дослідження зумовлена стрімким поширенням RAG-систем у пошукових і генеративних задачах, де якість відповіді залежить як від релевантності відібраного контексту, так і від коректності його використання генеративною мовною моделлю. Метою дослідження є огляд сучасних фреймворків оцінювання RAG-систем та метрик і проведення експериментальної перевірки впливу якості вибірки на показники генерації. У роботі проведено аналіз наукових публікацій, порівняння самих фреймворків оцінювання, машинний експеримент на основі систем векторного пошуку з подальшою генерацією відповіді. Для оцінки впливу фільтрації на якість вибірки та формування контексту виконано порівняння стандартного векторного пошуку та пошуку з попередньою фільтрацією. Отримані результати підтверджують, що оцінювання систем RAG має враховувати як метрики вибірки, так і метрики генерації, оскільки збільшення контексту без зменшення шуму не гарантує покращення якості відповіді.

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

Постановка проблеми. Системи Retrieval-Augmented Generation (RAG) поєднують пошук за зовнішніми джерелами з генерацією тексту великими мовними моделями (LLM), що дозволяє знижувати галюцинації та опиратися на актуальні і релевантні для домену дані [1]. Такий підхід широко використовується в різноманітних комп'ютерних системах та компонентах (модулях, сервісах, та ін.,) – корпоративних асистентах знань, підтримці клієнтів, пошуку по технічній документації та інших доменах, де відповідь має ґрунтуватися на заданому, наперед визначеному наборі даних (датасеті).

На сьогодні існує значна кількість підходів до вимірювання ефективності RAG-систем – від класичних метрик інформаційного пошуку до популярних фреймворків на кшталт RAGAS чи ARES та кастомізованих бенчмарків з власними метриками [2, 4, 5]. В роботах [6, 7] відзначається обмеження існуючих підходів та відсутність єдиного стандарту оцінки RAG-систем, оскільки різні бенчмарки використовують різні набори даних, окремі критерії релевантності вибірки та правильності відповіді, що ускладнює порівняння результатів. У такій ситуації систематичний огляд метрик і фреймворків разом із практичною перевіркою на реальній системі дає змогу упорядкувати знання, проаналізувати та проілюструвати застосування основних показників (метрик) на конкретних даних.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Мета дослідження – провести огляд фреймворків та метрик для оцінки RAG-систем, проаналізувати їхні сильні та слабкі сторони і підкріпити висновки експериментами на існуючій системі векторного пошуку з подальшим етапом генерації.

Серед основних, класичних, метрик оцінювання RAG-систем виділяють Precision@K, Recall@K, MRR, retrieval latency, а також метрики генерації Faithfulness і Answer Relevancy, що використовуються у фреймворку RAGAS [2, 3].

Аналіз останніх досліджень і публікацій. У дослідженнях RAG-систем застосовуються як reference-based, так і reference-free підходи до оцінювання якості відповіді та контексту [2, 3, 6]. Класичні метрики дають чітку кількісну оцінку якості вибірки, але вимагають наявності еталонної множини релевантних документів (reference-based) для кожного тестового запиту. Така множина зазвичай створюється вручну експертами або за допомогою дорогого процесу анотації, що обмежує масштаб і здатність до оновлення бенчмарків [6].

Підходи, де відсутня еталонна розмітка (так звані reference-free фреймворки), як правило покладаються на LLM, оскільки оцінюють релевантність контексту та коректність відповіді без явної розмітки, використовуючи LLM для виділення тверджень, порівняння з контекстом або генерації гіпотетичних питань [2]. Сильною стороною такого підходу є можливість швидко оцінювати нові пайплайни без попередньої анотації. До недоліків такого підходу можна віднести залежність результатів від вибору LLM, її калібровки та стабільності при запусках. Альтернативою може стати використання невеликих моделей, спеціально адаптованих та натренованих на синтетичних або частково розмічених даних для оцінки окремих компонентів RAG – LLM-суддей (LLM-as-judge). Надання LLM-судді високоякісних еталонних відповідей для навчання істотно покращує результати навіть для менш потужних моделей-суддів [14]. Сильні сторони таких тонко налаштованих суддів – передбачувана поведінка, нижча вартість та можливість адаптації до домену, а слабкі – залежність від наявності даних для навчання та ризики застосування в інших доменах.

Нехай q – текстовий запит користувача, D – набір документів (або їх фрагментів), проіндексований для пошуку. Множина документів, релевантних запиту q згідно з еталонною розміткою (ground truth), позначається як $R_q \subseteq D$. Результат вибірки для запиту q при обмеженні на топ-K документів – множина $T_k(q) \subseteq D$, тобто K документів, повернутих системою як найбільш релевантні (наприклад, за косинусною схожістю векторів текстового запиту та проіндексованих документів) [1, 6].

Типова схема RAG складається з таких послідовних кроків:

- введення запиту q користувачем;
- перетворення введеного запиту q на векторне представлення (embedding);
- пошук у векторній базі та формування контексту з топ-K найбільш релевантних документів $T_k(q)$;
- передача запиту q та отриманих документів $T_k(q)$ в LLM у якості контексту;
- генерація відповіді.

Можна сказати, що ефективність такої RAG-системи залежить як від того, наскільки $T_k(q)$ перетинається з R_q, так і від того, наскільки згенерована відповідь відповідає отриманому контексту та релевантна запиту користувача [2, 3].

Метрики якості вибірки оцінюють, наскільки добре система підбирає документи для подальшої передачі в LLM. Вони вимагають наявності еталонної множини релевантних документів R_q, для кожного запиту q з тестової множини запитів Q [6].

Precision@K визначає частку релевантних серед перших K результатів пошуку. Нехай $T_k(q)$ – множина топ-K документів, повернутих системою для запиту q, а R_q – множина релевантних документів для q. Тоді Precision@K:

$$Precision@K(q) = \frac{|R_q \cap T_k(q)|}{K} \quad (1)$$

Високе значення Precision@K свідчить про зменшення «шуму» у контексті та сприяє тому, щоб LLM спиралася на корисну інформацію [8].

Recall@K показує, яка частка всіх релевантних документів потрапила до топ-K результатів:

$$Recall@K(q) = \frac{|R_q \cap T_k(q)|}{|R_q|} \quad (2)$$

Якщо релевантних документів багато, то низьке значення (2) означає втрату важливої інформації для генерації [6]. Тобто можна сказати, що Precision@K оцінює якість, а Recall@K – повноту охоплення контексту.

Середній обернений ранг (Mean Reciprocal Rank, MRR) оцінює положення першого релевантного документу в ранжованому списку. Метрика є особливо корисною у сценаріях, де достатньо отримати один релевантний документ на початку списку [8].

Затримка вибірки (retrieval latency) – час від формулювання запиту до отримання контексту без урахування часу генерації відповіді. У практичних рекомендаціях зазначається, що затримка понад 500 мс може погіршувати досвід користувача, а добре оптимізоване векторне сховище зазвичай повертає результати за менш ніж 200 мс для середніх обсягів даних.

Варто відзначити, що reference-free фреймворки, як правило використовують похідні метрики – наприклад, Context Precision та Context Recall у RAGAS оцінюють релевантність та повноту контексту, переданого в LLM, на основі підтвердження фрагментів контексту. Метрика Context Entities Recall перевіряє покриття ключових сутностей з запиту у вибраному контексті. Noise Sensitivity оцінює, наскільки додавання нерелевантних документів до контексту погіршує якість відповіді [2, 4]. У бенчмарку MIRAGE введено інші метрики адаптивності RAG-систем: Noise Vulnerability, Context Acceptability, Context Insensitivity та Context Misinterpretation [4]. Фреймворк VERA об’єднує багатовимірні показники якості вибірки та генерації в єдиний узагальнений рейтинговий показник (ranking score) [10]. Бібліотека BERGEN стандартизує обчислення класичних метрик вибірки, зокрема Precision, Recall, MRR, а також метрик генерації (F1, Exact Match, ROUGE, BEM, LLMEval) для відтворюваних експериментів на множині тестових наборів даних [8].

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Faithfulness демонструє, наскільки кожне твердження (claim) у відповіді узгоджується з наданим контекстом. Вона має значення від 0 до 1, де вищі значення означають кращу узгодженість, а нижчі – наявність галюцинацій або використання зовнішніх знань моделі замість наданого контексту [9]. У RAGAS спочатку виділяють усі твердження у відповіді, потім для кожного перевіряють підтримку в контексті вручну або за допомогою LLM чи спеціалізованої моделі, зокрема HHEM-2.1-Open, після чого обчислюють частку підтверджених тверджень [2]. Інструмент ARES має схожу метрику Answer Faithfulness, але проводить оцінку за допомогою додатково адаптованих (навчених) на синтетичних даних легких суддів з подальшим уточненням через статистичне оцінювання, підсилене прогнозами (prediction-powered inference, PPI) [3].

Зазначимо, що метрики на кшталт Faithfulness перевіряють узгодженість відповіді з наданим контекстом, але не завжди враховують ситуації, коли контекст непридатний (Unanswerable), суперечливий (Inconsistent) або навмисно хибний (Counterfactual). Бенчмарк FaithEval показує, що навіть провідні моделі не завжди відхиляють такий контекст або відповідають «не знаю», а потужніші моделі не обов'язково краще справляються з вірністю контексту в цих умовах [11]. GaRAGe вводить метрику Relevance-Aware Factuality Score з анотаціями опорних фрагментів (grounding passages), аргументуючи це тим, що якісна оцінка RAG-систем повинна відповідати на питання не лише «чи відповідь узгоджена з контекстом», але й «чи контекст достатній і неупереджений» та «чи модель коректно обмежує відповідь доступною інформацією». Варто відмітити, що навіть сучасні моделі досягають на цій метриці значення не більше 60% і особливо слабшають на розрізнених джерелах даних, що залежать від часу [12].

Answer Relevancy оцінює, наскільки отриманий результат відповідає введеному запиту. Значення метрики коливається від 0 до 1, де вищі показники означають кращу відповідність запиту. Вона знижує оцінку для відповідей, які є неповними або містять зайву інформацію. У RAGAS це найчастіше реалізується через генерацію гіпотетичних питань до відповіді та порівняння їх із оригінальним запитом або через LLM-суддю [2]. ARES використовує окрему вимірювальну компоненту для релевантності відповіді [3].

Hallucination Rate – частка відповідей, що містять вигадану або непідтверджену контекстом інформацію, цей показник є оберненим до Faithfulness. Його вимірювання можливе за наявності анотованих наборів даних або за допомогою автоматичних детекторів. Інструмент RAGTruth надає великий набір відповідей RAG-систем з ручною анотацією галюцинацій на рівні випадків і слів, що дозволяє тренувати та перевіряти детектори галюцинацій і методи їх усунення [10]. Серед недоліків даного підходу – вартість створення та оновлення подібних датасетів для галюцинацій і доменна специфіка. За наявності еталонної розмітки, деякі метрики – BLEU, ROUGE, F1 за токенами або n-грамами – порівнюють згенеровану відповідь з еталонною. Такий підхід корисний в експериментах з фіксованим набором запитів, але менш інформативний для відкритих запитань [6].

Оцінка якості RAG-систем усереднює різні аспекти, при цьому окремі типи помилок генератора (наприклад, ігнорування контексту, вигадування фактів, неповна відповідь) можуть залишатися непоміченими, якщо використовувати лише сумарні метрики

або кореляцію з одним суддею. Бенчмарк GroUSE формалізує сім типів помилок і пропонує майже півтори сотні юніт-тестів для перевірки калібровки та дискримінаційної здатності суддів [14]. Автори показують, що наявні автоматизовані оцінювачі часто не виявляють важливі типи помилок навіть при використанні сучасних потужних LLM-моделей в якості судді. В такому випадку рекомендується доповнювати агреговані метрики та кореляційні порівняння юніт-тестами на конкретні типи помилок. Це узгоджується з потребою в більш деталізованій діагностиці метрик вибірки та генерації [6, 7].

Окремо варто зазначити, що стандартні метрики вибірки і генерації розроблені переважно для невеликих контекстів і коротких відповідей. Для RAG-систем із довгими відповідями (long-form RAG), зокрема у завданнях підсумування великих документів з посиланнями на джерела, потрібні додаткові показники. SummHay пропонує додаткові метрики Coverage (наскільки відповідь охоплює ключові змісти з контексту) та Citation Accuracy (точність посилань на фрагменти документів) [13]. Існують і інші підходи – наприклад, CRUX фокусуються на контрольованій області контексту та оцінці покриття на основі запитань [15]. Це підкреслює, що поточна генерація та оцінка long-form RAG залишаються відкритою проблемою і вимагають спеціалізованих метрик та бенчмарків.

Через відсутність єдиного підходу до оцінки якості RAG-систем, також пропонуються уніфіковані процеси та систематична оцінка метрик вибірки, генерації і додаткових вимог (наприклад, RGAR-подібні схеми). Бібліотека BERGEN спрямована на підвищення відтворюваності, оскільки фіксує набори даних, моделі пошуку (retrievers), моделі повторного ранжування (rerankers), LLM і набір метрик, що дає змогу повторювати та порівнювати експерименти за однакових умов [8]. Сильна сторона таких ініціатив – зниження фрагментації, а слабка – необхідність широкого прийняття спільнотою та підтримки оновлень бенчмарків і метрик.

Додаткові бенчмарки та набори даних, зазначені вище, включають KILT (об’єднані knowledge-intensive задачі на єдиному Wikipedia dump), BEIR (гетерогенний бенчмарк інформаційного пошуку), а також спеціалізовані фреймворки для long-form RAG (CRUX, Long²RAG з метрикою Key Point Recall) та оцінки детекторів галюцинацій (ORION, LUMINA) [6, 7]. У табл. 1 наведено основні фреймворки та бенчмарки для оцінки RAG-систем за типами метрик та умовами їх застосування.

Підсумовуючи вищесказане, сильні сторони сучасних підходів полягають у розмаїтті метрик вибірки і генерації, у зручності оцінки без еталонної розмітки для швидких ітерацій та у появі бенчмарків для окремих аспектів (галюцинації, вірність контексту, довгий контекст) оцінки RAG. Слабкі ж сторони включають залежність методів без еталонної розмітки від калібровки та стабільності LLM-судді, обмеження застосування поточних RAG-систем для довгих відповідей та відсутність єдиного стандарту оцінювання.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Таблиця 1

Фреймворків і бенчмарків оцінки RAG-систем

Фреймворк / бенчмарк	Особливість	Метрики вибірки	Метрики генерації
RAGAS	відсутність еталонної розмітки, використання LLM-суддів	Context Precision, Context Recall, Context Entities Recall, Noise Sensitivity	Faithfulness, Answer Relevance
ARES	використання «легких» суддів + PPI	Context Relevance	Answer Faithfulness, Answer Relevance
VERA	валідація та ранжування через об'єднання метрик; наявність bootstrap-статистики	Retrieval Precision, Retrieval Recall	Faithfulness, Answer Relevance
MIRAGE	наявність компактного, але складного тестового датасету	Precision, Recall, NDCG (Normalized Discounted Cumulative Gain), Accuracy	Noise Vulnerability, Context Acceptability, Context Insensitivity, Context Misinterpretation
RAGBench + TRACe	5 прикладних доменів, повна анотація контексту, «пояснювальність»	Relevance (R)	Utilization (T), Adherence (A), Completeness (C)
BERGEN	бібліотека для стандартизації експериментів та метрик, відтворюваності	Precision, Recall, MRR (Mean Reciprocal Rank), NDCG (Normalized Discounted Cumulative Gain)	F1, Exact Match (EM), ROUGE, BEM, LLMEval (LLM-based Answer Equivalence)
FaithEval	бенчмарк для перевірки вірності відповіді моделі наданому контексту	-	вірність при Unanswerable / Inconsistent / Counterfactual контексті
GaRAGe	бенчмарк з анотаціями опорних фрагментів та «людськими» відповідями	Grounding Quality (якість набору опорних фрагментів), Pass_r (частка релевантних фрагментів у контексті)	Unadjusted Factuality Score, Eligibility Score, Unadjusted Relevance-Aware Factuality, Relevance-Aware Factuality, Factuality Score, Deflection Score, Attribution Score
RAGTruth	оцінювання та детекція галюцинацій з тонкою розміткою як на рівні всієї відповіді, так і на рівні окремих слів	-	Hallucination Frequency, Hallucination Density, Span-level Precision, Span-level Recall, Case-level Accuracy
SummHay (Summary of a Haystack)	бенчмарк для довгого контексту	-	Coverage, Citation Accuracy, Summarization Quality
GroUSE	бенчмарк для мета-оцінки суддів	-	7 режимів збою генератора, юніт-тести для суддів

Викладення основного матеріалу дослідження. Експериментальна частина побудована на системі векторного пошуку у GCP Firestore. Використовується датасет із 200 резюме кандидатів різноманітних IT-спеціалізацій. З метою екстракції тексту, генерації векторних представлень та виокремлення структурованих метаданих кожен документ було автоматично опрацьовано з використанням моделей сервісу Google Vertex AI. До метаданих відносяться посада, країна, володіння мовами, досвід, навички та рівень кваліфікації. Оброблені дані збережені в колекції Firestore у структурованому вигляді (текст, векторне представлення, метадані), що забезпечує можливість застосування стандартного векторного пошуку та пошуку з попередньою фільтрацією за метаданими у межах єдиної колекції. Для оцінки якості вибірки використовується набір з 10 тестових запитів (табл. 2), кожен запит задається лише текстом.

Таблиця 2

Тестові запити

№ з/п	Запит	К-ть фільтрів
1	Business Analyst from Poland	1
2	Frontend Developer with 8+ years experience	1
3	Senior DevOps Engineer from Ukraine	2
4	QA Engineer who speaks German	1
5	BA from Poland with 5+ years experience	2
6	Fullstack Developer from Poland who speaks English	2
7	Backend Developer with Python and Docker experience	2
8	Senior Backend Developer with 10+ years experience who speaks German	3
9	Cloud Architect from Poland or Germany with 12+ years experience and Kubernetes skills	3
10	Senior Business Analyst with 10+ years experience, speaking German, skilled in JIRA and Confluence	4

Як було згадано вище, метрики Precision@K, Recall@K та MRR вимагають еталонної множини релевантних документів для кожного тестового запиту. В даній роботі релевантність документу визначається семантикою запиту (наприклад, для запиту «Business Analyst from Poland» релевантними є резюме бізнес-аналітиків з Польщі, а не довільні документи з Польщі). Множина релевантних документів отримана експертною розміткою. Допустимо вважати релевантними резюме вищого рівня кваліфікації, ніж задано в запиті (наприклад, Senior замість Middle або Lead замість Senior).

Для оцінки якості генерації проведено компактний експеримент у форматі повного RAG-пайплайну. Для кожного тестового запиту виконується векторний пошук, релевантні документи об’єднуються в один текстовий контекст, який разом із запитом передається в LLM. Генерація відповідей виконується однією моделлю (далі модель-генератор), оцінка метрик Faithfulness та Answer Relevancy – окремою моделлю (далі модель-суддя). Застосування окремих моделей для генератора та судді зменшує ризик порожніх або заблокованих відповідей судді через довгий контекст або обмеження безпеки, оскільки генератор працює з повним контекстом, а суддя отримує скорочені промпти з обмеженою довжиною контексту та відповіді. Якість згенерованих відповідей

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

оцінюється за двома метриками, близькими до описаних в RAGAS, а саме Faithfulness (частка тверджень у відповіді, підтверджених контекстом) та Answer Relevancy (наскільки відповідь релевантна запиту). Цей експеримент має обмежений обсяг (10 запитів, K = 3, 5, 10, один генератор, один суддя) і призначений виключно для ілюстрації застосування метрик генерації на тій самій RAG-системі, що використовується і для оцінки метрик вибірки.

Нижче наведено значення метрик вибірки по кожному запиту окремо із застосуванням різних типів пошуку – табл. 3 для стандартного векторного пошуку, табл. 4 для пошуку з попередньою фільтрацією. Метрики Precision@5, Recall@5 та MMR описані вище. Затримка вибірки (мс) для кожного запиту – це середнє значення по 1000 вимірів. При такому обсязі середнє значення має приблизно нормальний розподіл, і 95% довірчий інтервал для математичного сподівання будується як t-інтервал Стьюдента – цей підхід не вимагає знання справжньої дисперсії і дає коректне покриття при скінченній вибірці [16].

Таблиця 3

Метрики по запитам, стандартний пошук

Запит	Precision@5	Recall@5	MMR	Затримка вибірки (мс)
№1	0,60	1,00	1,00	43,51±0,27
№2	0,80	0,36	1,00	35,88±0,31
№3	0,60	0,27	1,00	64,26±0,33
№4	1,00	0,50	1,00	58,65±0,47
№5	0,60	1,00	1,00	61,95±0,45
№6	0,20	1,00	0,50	57,88±0,28
№7	0,80	0,40	1,00	25,52±0,16
№8	0,40	0,50	1,00	58,82±0,62
№9	0,40	1,00	1,00	41,70±0,24
№10	0,20	1,00	1,00	40,91±0,45

Таблиця 4

Метрики по запитам, пошук з фільтрацією

Запит	Precision@5	Recall@5	MMR	Затримка вибірки (мс)
№1	0,60	1,00	1,00	19,40±0,13
№2	1,00	0,45	1,00	19,88±0,14
№3	1,00	0,45	1,00	20,82±0,17
№4	1,00	0,50	1,00	19,79±0,23
№5	0,60	1,00	1,00	18,59±0,14
№6	1,00	1,00	1,00	17,71±0,15
№7	0,80	0,40	1,00	27,48±0,33
№8	1,00	1,00	1,00	17,73±0,22
№9	0,40	1,00	1,00	13,64±0,20
№10	0,20	0,10	0,00	17,03±0,21

В табл. 5 наведено середні значення метрик якості (для різних значень K) та затримки вибірки для обох типів пошуку.

Таблиця 5

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

Можна зробити висновок, що якість видачі залежить від складності запиту та розміру еталонної розмітки, а високе значення MRR свідчить про те, що перший релевантний документ зазвичай потрапляє на перші позиції видачі. Зауважимо, що пошук з попередньою фільтрацією значно швидший за стандартний пошук завдяки істотному зменшенню обсягу даних для виконання операції векторного пошуку. На рис. 1. зображені метрики якості вибірки (Precision@K та Recall@K), при K = 3, 5, 10).

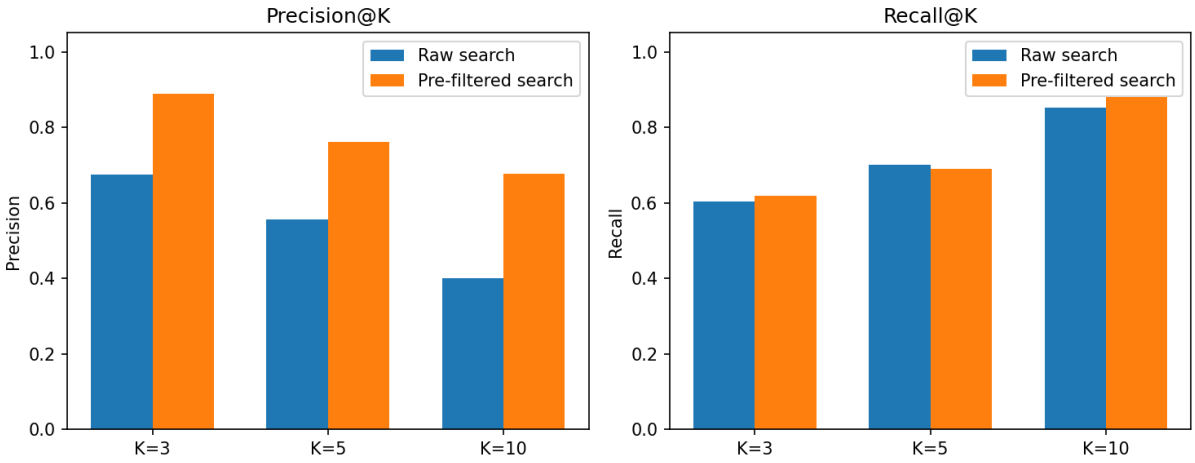

Рисунок 1 - Метрики якості вибірки для стандартного пошуку та пошуку з попередньою фільтрацією

Пошук з фільтрацією дає вищі значення Precision за рахунок зменшення шуму в контексті, а значення Recall може змінюватись залежно від того, чи не відсікають фільтри частину релевантних документів. Для оцінки якості генерації виконано компактний експеримент, в якому кожен з 10 тестових запитів оброблено через RAG-пайплайн, після чого модель-суддя оцінила метрики Faithfulness та Answer Relevancy за схемою RAGAS [2]. Експеримент проведено для стандартного векторного пошуку та пошуку з попередньою фільтрацією при різній величині контексту (K=3 і K=5). Для генерації ві-

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

дповідей використовувалась модель Gemini 2.5 Flash. Для оцінки значень метрик генерації застосовувалась окрема модель-суддя Gemini 2.5 Pro. Результат представлено в табл. 6.

Таблиця 6

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

Як бачимо, середні значення Faithfulness у випадку стандартного векторного пошуку дорівнюють 1,00 при K=3 та 0,72 при K=5. Оскільки при K=5 більше документів у контексті, відповіді можуть містити більше тверджень, частину з яких модель-суддя не вважає повністю підтвердженою, через що і знижується значення метрики. Для пошуку з попередньою фільтрацією значення Faithfulness відповідно 1,00 (K=3) та 0,94 (K=5) – при K=5 цей показник вищий, ніж у стандартного пошуку за рахунок більш релевантного контексту після фільтрації метаданих. Середні значення метрики Answer Relevancy для стандартного пошуку дорівнюють 0,94 (K=3) та 0,65 (K=5). Як бачимо, при K=5 середня релевантність нижча, що може бути пов’язано з більш «розмитими» відповідями при більшому контексті. Для пошуку з попередньою фільтрацією ці значення 0,78 (K=3) та 0,80 (K=5) відповідно. Табл. 7 демонструє метрики Faithfulness та Answer Relevancy для обох типів пошуку по кожному запиту окремо (K=5).

Таблиця 7

Метрики генерації по запитам

Запит	Faithfulness (стандартний)	Answer Relevancy (стандартний)	Faithfulness (з фільтрацією)	Answer Relevancy (з фільтрацією)
№1	1,00	0,90	1,00	1,00
№2	0,75	0,10	0,80	1,00
№3	0,75	1,00	1,00	1,00
№4	0,00	0,10	1,00	0,10
№5	0,67	0,20	1,00	0,20
№6	1,00	1,00	1,00	0,90
№7	0,60	1,00	0,60	1,00
№8	0,50	0,90	1,00	1,00
№9	1,00	0,50	1,00	1,00
№10	1,00	1,00	1,00	1,00

Якщо проаналізувати отримані значення окремо для кожного з тестових запитів, то варто відмітити результат для query_3 (запит «QA Engineer who speaks German»). Для стандартного пошуку (при K=5) модель дала дуже коротку відповідь і суддя поста-

вив низькі значення Faithfulness та Answer Relevancy (через те, що в тексті чанку не згадано мову явно). Для пошуку з попередньою фільтрацією (при K=5) ми маємо інші значення (Faithfulness 1,00, Answer Relevancy 0,10) – тобто твердження вже підтверджені контекстом, але релевантність оцінена низько через дуже стислу відповідь. Такі відмінності оцінки по типах пошуку коректно відображають якість згенерованої відповіді та впливають на середнє по 10 запитах.

Порівняння метрик вибірки двох підходів до векторного пошуку показує, що попередня фільтрація документів за метаданими (країна, посада, мови, досвід, навички тощо) суттєво підвищує Precision при будь-яких значеннях K, а значення Recall та MRR залишаються майже незмінними, тобто фільтри не відсікають критичну кількість релевантних документів на даному наборі запитів. Затримка в формуванні вибірки також знижується через зменшення обсягу даних для виконання операції векторного пошуку. Попередня фільтрація за метаданими також позитивно впливає на якість згенерованих відповідей при більшому K, коли контекст довший і ризик «шумових» документів вищий. Щодо стандартного векторного пошуку – зі збільшенням значення K спостерігається зниження Faithfulness та Answer Relevancy, що очікувано – при більшому контексті відповіді можуть містити більше тверджень або ставати «розмитими».

У ході експериментів виявлено, що LLM-суддя за схемою RAGAS дає інтерпретовані оцінки Faithfulness та Answer Relevancy, але їх надійність обмежена варіативністю судді та залежністю від обсягу переданого контексту. У нашому експерименті оцінки судді не порівнювались з експертною анотацією, тому можна інтерпретувати лише зміну метрик при зміні вибірки, а не їх абсолютні значення. Важливо також зазначити, що передача судді занадто короткого контексту (перших 1000 символів) інколи призводила до помилково низьких Faithfulness для окремих запитів – це технічна, але критична умова валідності оцінки підходу без еталонної розмітки, яка майже не згадується в описах RAGAS-подібних пайплайнів. Збільшення ліміту до 10000 символів усунуло проблему.

Серед обмежень дослідження можна відзначити, що обсяг тестової вибірки становить лише десять запитів, тому довірчі інтервали залишаються широкими. Поточна вибірка достатня для ілюстрації ефектів і порівняння методів, але не для прецизійної оцінки популяційних параметрів. Збільшення набору до 50-100 запитів дало б точніші оцінки середніх та вужчі довірчі інтервали. Експеримент з генерації виконано з однією моделлю генератора та одним суддею. Інші моделі можуть давати різні співвідношення Faithfulness та Answer Relevancy при тій самій якості вибірки. Дослідження проводилось на одному датасеті (200 документів) та одній інфраструктурі (GCP Firestore, Vertex AI). Тестування на інших масштабах (тисячі документів, інші векторні бази) та інших реалізаціях RAG залишається емпіричним питанням.

Висновки. У роботі проведено огляд поширених фреймворків і метрик оцінювання RAG-систем та експериментальну перевірку ключових рекомендацій на датасеті з векторним пошуком у Firestore та генерацією відповідей LLM. На основі проведеного експерименту підтверджено, що пошук з попередньою фільтрацією значно підвищує

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Precision вибірки при практично незмінному Recall і MRR на датасеті зі структурованими метаданими. Отриманий контекст пошуку з попередньою фільтрацією покращує значення Faithfulness та Answer Relevancy, що узгоджується з постулатом про залежність якості генерації від чистоти контексту. При попередній фільтрації затримка вибірки суттєво знижується завдяки зменшенню обсягу даних для векторного пошуку. Недостатній обсяг контексту, переданого судді, призводить до помилково низьких оцінок Faithfulness. Збільшення контексту (значення K) покращує Recall, але при стандартному пошуку може погіршувати Faithfulness та Answer Relevancy. Пошук з попередньою фільтрацією дозволяє використовувати більше значення K без пропорційного погіршення метрик генерації.

ЛІТЕРАТУРА / REFERENCES

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. Park Chanhee, Moon H., Park Chanjun, Lim H. MIRAGE: A Metric-Intensive Benchmark for Retrieval-Augmented Generation Evaluation. Computer Science. Computation and Language. – 2025. DOI: 10.48550/arXiv.2504.17137

5. Friel R., Belyi M., Sanyal A. RAGBench: Explainable Benchmark for Retrieval-Augmented Generation Systems. Computer Science. Computation and Language. – 2024. DOI: 10.48550/arXiv.2407.11005

6. 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

7. Gan A., Yu H., Zhang K., Liu Q., Yan W., Huang Z., Tong S., Hu G. Retrieval Augmented Generation Evaluation in the Era of Large Language Models: A Comprehensive Survey. Computer Science. Computation and Language. – 2025. DOI: 10.48550/arXiv.2504.14891

8. Rau D., Déjean H., Chirkova N, Formal T., Wang S., Nikoulina V., Clinchant S. BERGEN: A Benchmarking Library for Retrieval-Augmented Generation. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), Findings. – 2024. P. 5897-5913. DOI: 10.48550/arXiv.2407.01102

9. Niu C., Wu Y., Zhu J., Xu S., Shum K., Zhong R., Song J., Zhang T. RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL). 2024. P. 10794-10817. DOI: 10.48550/arXiv.2401.00396

10. Ding T., Banerjee A., Mombaerts L., Li Y., Borogovac T., Weinstein J. P. VERA: Validation and Evaluation of Retrieval-Augmented Systems. Computer Science. Information Retrieval. – 2024. DOI: 10.48550/arXiv.2409.03759
11. Ming Y., Purushwalkam S., Pandit S., Ke Z., Nguyen X., Xiong C., Joty S. FaithEval: Can Your Language Model Stay Faithful to Context, Even If "The Moon is Made of Marshmallows". Computer Science. Computation and Language. – 2024. DOI: 10.48550/arXiv.2410.03727
12. Sorodoc I.-T., Ribeiro L., Blloshmi R., Davis C., de Gispert A. GaRAGe: A Benchmark with Grounding Annotations for RAG Evaluation. Computer Science. Computation and Language. – 2025. DOI: 10.48550/arXiv.2506.07671
13. Laban P., Fabbri A. R., Xiong C., Wu C.-S. Summary of a Haystack: A Challenge to Long-Context LLMs and RAG Systems. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). – 2024. DOI: 10.48550/arXiv.2407.01370
14. Krumdick M., Lovering C., Reddy V., Ebner S., Tanner C. No Free Labels: Limitations of LLM-as-a-Judge Without Human Grounding. Computer Science. Computation and Language. – 2025. DOI: 10.48550/arXiv.2503.05061
15. Ju J.-H., Verberne S., de Rijke M., Yates A. Controlled Retrieval-augmented Context Evaluation for Long-form RAG. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). – 2025. DOI: 10.48550/arXiv.2506.20051
16. Casella G., Berger R. L. Statistical Inference. 2nd ed. Pacific Grove: Duxbury, 2002.

Received 06.04.2026.
Accepted 08.04.2026.
Published 30.04.2026

A comprehensive survey of modern frameworks and evaluation metrics for RAG systems

The relevance of the study is driven by the rapid proliferation of RAG systems in search and generative tasks, where response quality depends on both the relevance of the retrieved context and the correctness of its utilization by generative language model. The objective of the research is to review modern metrics and frameworks for evaluating RAG systems and experimentally verify the impact of retrieval quality on generation metrics. The study analyzes scientific publications, compares evaluation frameworks, and conducts a machine experiment using a vector search system followed by response generation. To evaluate the impact of filtering on retrieval quality and context formation, we compare standard vector search with pre-filtered search. The obtained results confirm that RAG system evaluation must account for both retrieval and generation metrics, as increasing context size without reducing noise does not guarantee improved response quality.

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

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Клименко Іван Вікторович – доцент, к.е.н., доцент кафедри комп'ютерних і інформаційних технологій Українського державного університету науки і технологій.
ORCID: https://orcid.org/0000-0001-5149-3974
Лебідь Євген Андрійович – аспірант кафедри комп'ютерних і інформаційних технологій Українського державного університету науки і технологій.
ORCID: https://orcid.org/0009-0007-4277-2083

Klymenko Ivan – Associate Professor, Candidate of Economics Sciences, Associate Professor of the Department of Computer Information Technologies, Ukrainian State University of Science and Technology.
ORCID: https://orcid.org/0000-0001-5149-3974
Lebid Yevhen – PhD Student of the Department of Computer Information Technologies, Ukrainian State University of Science and Technology.
ORCID: https://orcid.org/0009-0007-4277-2083