УДК 004.8:004.912

ОЦІНЮВАННЯ РІШЕНЬ RAG-СИСТЕМ ЩОДО ВІДПОВІДІ АБО ВІДМОВИ З УРАХУВАННЯМ ДОСТАТНОСТІ ДОКАЗІВ

Вербицький О.С.[1]; Чернова Л.С.[2], д.т.н., проф.
[1,2]Національний університет кораблебудування імені адмірала Макарова
[1,2]Україна, Миколаїв
[1]ORCID: 0009-0000-6848-8880; [2]ORCID: 0000-0002-0666-0742

Анотація. *Запропоновано дворівневу модель оцінювання рішень RAG-систем щодо відповіді або відмови. Вона розрізняє наявність доказів у корпоративному корпусі та їх достатність в отриманому контексті. Визначено п'ять станів доказової підтримки, відповідні дії системи, метрики оцінювання й протокол пілотної перевірки. Емпіричну валідацію заплановано.*

Ключові слова: *інформаційні технології; управління IT-проєктами; комп'ютерні науки; достатність доказів; коректність відмови.*

Вступна частина. Системи генерації, доповненої пошуком (Retrieval-Augmented Generation, RAG), формують відповіді на підставі отриманого документального контексту. Проте наявність фрагментів не гарантує достатності доказів: контекст може повністю або частково підтверджувати запитувану інформацію, бути нерелевантним або містити суперечності. Тому важливо оцінювати не лише правильність відповіді, а й вибір дії системи: надати повну чи обмежену відповідь, повторити пошук, утриматися від відповіді або повідомити про конфлікт джерел. Фреймворк UAEval4RAG систематизує шість категорій запитів, на які неможливо відповісти в межах конкретної бази знань [1]. Joren et al. [2] розмежовують недостатність отриманого контексту та нездатність моделі використати наявний достатній контекст. Zhang і Wu [3] запропонували п'ятирівневу шкалу достатності доказів (benchmark): повна й часткова підтримка, нерелевантні та відсутні докази, суперечливі відомості. У цій роботі шкалу [3] поєднано з розмежуванням достатності контексту, описаним у [2], та адаптовано до корпоративного пошуку: стан L3 означає, що релевантні документи є в повному корпусі, але не потрапили до первинного контексту.

Метою дослідження є розроблення теоретико-методичної моделі оцінювання рішень RAG-систем щодо відповіді або відмови з урахуванням двох рівнів: наявності доказів у повному корпоративному корпусі та їх достатності в отриманому контексті.

Основна частина. Запропонована модель оцінювання охоплює два рівні. На корпусному рівні експертна розмітка встановлює наявність достатніх доказів і суперечностей у повній базі знань. На рівні пошуку перевіряють, чи потрапили релевантні фрагменти до контексту, переданого генератору. Для стану L3 передбачено до двох повторних пошуків. Такий поділ дає змогу відрізнити помилку пошуку від реальної відсутності даних у корпусі.

У табл. 1 узагальнено п'ять станів доказової підтримки з урахуванням повного корпусу та первинно отриманого контексту, а також еталонні дії RAG-системи.

Таблиця 1 - Стани доказової підтримки та рекомендовані дії RAG-системи

Стан	Характеристика доказів	Рекомендована дія системи
L1	Повний корпус і отриманий контекст містять достатні, несуперечливі докази.	Надати повну відповідь на підставі отриманих доказів.
L2	Документи підтверджують лише частину запитаної інформації; суперечностей немає.	Надати підтверджену частину й позначити обмеження відповіді.
L3	Потрібні документи є в повному корпусі, але відсутні в первинно отриманому контексті.	Повторити пошук; якщо достатніх доказів не отримано — утриматися від відповіді.
L4	Після перевірки повного корпусу підтверджень не виявлено.	Утриматися від відповіді та зазначити, що в перевіреному корпусі підтвердження не знайдено.
L5	Суперечливі документи є в корпусі; обидві версії наявні в отриманому контексті.	Повідомити про суперечність і подати обидві версії, не обираючи одну без правила пріоритетності.

Для L3 наявність потрібних документів у корпусі встановлюється експертною перевіркою: якщо повторний пошук не відновив контекст, система повинна утриматися від відповіді. Стан L5 застосовується, коли обидві суперечливі версії доступні в отриманому контексті; якщо одна з них не потрапила до контексту, це випадок L3.

Показники оцінювання та протокол подальшої перевірки. Для кількісного оцінювання визначено п’ять показників: частку правильних рішень (Correct Decision Rate, CDR), частку відновлення пошуку (Retrieval Recovery Rate, RRR), частку необґрунтованих відповідей (Unsupported Answer Rate, UAR), частку хибних повних відмов (False Refusal Rate, FRR) і частку коректного розпізнавання конфліктів (Conflict Recognition Rate, CRR).

CDR = 100 × (кількість запитів, для яких виконано еталонний протокол дій / загальна кількість запитів);

RRR = 100 × (кількість випадків L3, для яких повторний пошук відновив достатній контекст / загальна кількість випадків L3);

UAR = 100 × (кількість необґрунтованих однозначних відповідей у нерозв’язаних випадках L3 та L4 / загальна кількість нерозв’язаних випадків L3 та L4);

FRR = 100 × (кількість повних відмов для станів L1–L2 / загальна кількість запитів станів L1–L2);

CRR = 100 × (кількість коректно опрацьованих випадків L5 / загальна кількість випадків L5).

Показники узгоджено з еталонними діями табл. 1: CDR фіксує дотримання протоколу дій, RRR – відновлення контексту для стану L3, UAR – необґрунтовані відповіді там, де відповідати не можна, FRR – повні відмови для станів L1 і L2 (для L2 повна відмова зараховується як хибна), CRR – коректне опрацювання суперечностей.

Пілотну перевірку заплановано на 100 запитах – по 20 на кожен стан. Два експерти незалежно визначатимуть мітки; розбіжності буде узгоджено, початкову узгодженість оцінено за κ Коена. Для L3 потрібні документи є в корпусі, але відсутні в первинному контексті; для L4 їх не виявлено в повному корпусі; для L5 обидві суперечливі версії доступні в контексті. Порівняють стандартну RAG-систему та RAG із перевіркою достатності доказів і до двох повторних пошуків; решта параметрів буде однаковою. Відповіді оцінюватимуть за категоріями F/P/U/C та показником ESS із попереднього дослідження. Вибірка має пілотний характер, її обсяг надалі буде уточнено. Запропонована модель стосується комп'ютерних наук та інформаційних технологій; в управлінні IT-проєктами її можна застосовувати для приймального тестування RAG-компонентів, зокрема для оцінювання достатності доказів і коректності відмов.

Висновки. Запропоновано дворівневу модель, що розрізняє наявність доказів у повному корпусі та їх достатність у первинному контексті. П'ять станів L1–L5 пов'язано з еталонними діями та показниками CDR, RRR, UAR, FRR і CRR. Визначено протокол пілотної перевірки; емпіричну валідацію буде проведено на наступному етапі.

Використання інструментів ШІ. Під час підготовки цієї роботи автори використовували Grammarly для перевірки граматики та орфографії. Після використання цього інструменту автори переглянули та за потреби відредагували текст і несуть повну відповідальність за зміст публікації.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] Peng, X., Choubey, P. K., Xiong, C., & Wu, C.-S. (2025). Unanswerability evaluation for retrieval augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 8452–8472). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.415

[2] Joren, H., Zhang, J., Ferng, C.-S., Juan, D.-C., Taly, A., & Rashtchian, C. (2025). Sufficient context: A new lens on retrieval augmented generation systems. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR 2025)*. https://proceedings.iclr.cc/paper_files/paper/2025/file/33dffa2e3d2ab74a783d1a8c292f66d9-Paper-Conference.pdf

[3] Zhang, H., & Wu, W. (2026). Do LLMs know when evidence is insufficient? An evidence sufficiency benchmark for answer-abstention calibration in retrieval-augmented generation. *Computers, Materials & Continua, 89*(1), 69. https://doi.org/10.32604/cmc.2026.086343

Verbytskyi O. S., Chernova L. S.

Assessing rag systems answering and abstention decisions based on evidence sufficiency

Abstract. *The paper proposes a two-level model for assessing RAG systems' decisions to answer or abstain from answering. The model separates evidence availability in the full corporate corpus from evidence sufficiency in the retrieved context. It defines five evidence states with corresponding system actions, five evaluation metrics, and a pilot validation protocol. The model can support IT project management, in particular acceptance testing of RAG components; empirical validation is planned for the next stage.*

Keywords: *information technologies; IT project management; computer science; evidence sufficiency; correct abstention.*